LEGAL KNOWLEDGE AND INFORMATION SYSTEMS

Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, doctoral dissertations, textbooks, handbooks and proceedings volumes.
The FAIA series contains several sub-series, including 'Information Modelling and Knowledge Bases' and 'Knowledge-Based Intelligent Engineering Systems'. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. An editorial panel of internationally well-known scholars is appointed to provide a high quality selection.

Volume 302

Recently published in this series

Vol. 301. V. Sornlertlamvanich, P. Chawakitchareon, A. Hansuebsai, C. Koopipat, B. Thalheim, Y. Kiyoki, H. Jaakkola and N. Yoshida (Eds.), Information Modelling and Knowledge Bases XXIX
Vol. 300. I. Aguiló, R. Alquézar, C. Angulo, A. Ortiz and J. Torrens (Eds.), Recent Advances in Artificial Intelligence Research and Development – Proceedings of the 20th International Conference of the Catalan Association for Artificial Intelligence, Deltebre, Terres de l'Ebre, Spain, October 25–27, 2017
Vol. 299. A.J. Tallón-Ballesteros and K. Li (Eds.), Fuzzy Systems and Data Mining III – Proceedings of FSDM 2017
Vol. 298. A. Aztiria, J.C. Augusto and A. Orlandini (Eds.), State of the Art in AI Applied to Ambient Intelligence
Vol. 297. H. Fujita, A. Selamat and S. Omatu (Eds.), New Trends in Intelligent Software Methodologies, Tools and Techniques – Proceedings of the 16th International Conference (SoMeT_17)
Vol. 296. V.E.Balas, L.C. Jain, X. Zhao and F. Shi (Eds.), Information Technology and Intelligent Transportation Systems – Proceedings of the 2nd International Conference on Information Technology and Intelligent Transportation Systems (ITITS 2017), Xi'an, China, June 10, 2017
Vol. 295. J. Mizera-Pietraszko, R. Rodriguez Jorge, D.M. Almazo Pérez and P. Pichappan (Eds.), Advances in Digital Technologies – Proceedings of the 8th International Conference on Applications of Digital Information and Web Technologies ICADIWT 2017

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

Legal Knowledge and Information Systems

JURIX 2017: The Thirtieth Annual Conference

Edited by

Adam Wyner

University of Aberdeen, UK

and

Giovanni Casini

University of Luxembourg, Luxembourg

IOS Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-61499-837-2 (print)
ISBN 978-1-61499-838-9 (online)
Library of Congress Control Number: 2017962232

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

Preface

We are pleased to present to you the proceedings of the 30th International Conference on Legal Knowledge and Information Systems – JURIX 2017. For three decades, the JURIX conferences have been held under the auspices of the Dutch Foundation for Legal Knowledge Based Systems (www.jurix.nl). In the time, it has become a European conference in terms of the diverse venues throughout Europe and the nationalities of participants. The conference continues to address familiar topics and extending known techniques as well as reaching out to newer topics such as question-answering and using data-mining and machine-learning.

The 2017 edition of JURIX, which runs from 13–15 December, takes place in Luxembourg City, Luxembourg, on the Kirchberg Campus of the University of Luxembourg. We received 42 submissions for this edition, 12 of which were selected for publication as full papers (10 pages in the proceedings) and 13 as short papers (six pages in the proceedings), for an acceptance rate of around 59%. All papers were rigorously reviewed. The strongest papers were accepted as full-papers, for an acceptance rate of 28.6%, while borderline or weakly acceptable papers were accepted as short papers, making up 30% of accepted papers. The papers address a wide range of topics in Artificial Intelligence and Law, such as argumentation, norms, evidence, belief revision, citations, case based reasoning, and ontologies; diverse techniques were applied such as information retrieval and extraction, machine learning, semantic web, and network analysis amongst others; the textual sources included legal cases, Bar Examinations, and legislative/regulatory documents.

This year, our invited speakers lead AI and Law research and development in industry and government. One speaker was Tonya Custis, who is a Research Director at Thomson Reuters, where she leads a team of Research Scientists performing applied research in Artificial Intelligence technologies. She is currently leading projects that explore Question Answering and Natural Language Understanding in the Legal domain. Our other speaker was Monica Palmirani, who is a professor in Computer Science and Law and Legal Informatics at University of Bologna, School of Law. Amongst other activities, she has been a lead on efforts to develop the OASIS standards LegalDocML and the LegalRuleML, which aim to make the structure and content of legal documents machine-readable. Our speakers highlight the impact of current work of Artificial Intelligence and Law on real work practice.

In addition to the main conference program, the workshops added opportunities for work focussed on research beyond the usual JURIX scope. The First Workshop on Technologies for Regulatory Compliance provided a forum for discussion of research on technologies for regulatory compliance which use semantic resources or Artificial Intelligence techniques. The Fourth Workshop on Legal Data Analysis of the Central European Institute of Legal Informatics (LDA: CEILI) focussed on the representation, analysis, and reasoning with legal data in information systems from a lawyer's and citizen's perspective. The Ninth Workshop on Artificial Intelligence and the Complexity of Legal Systems (AICOL) welcomed research in AI, political and legal theory, jurisprudence, philosophy of technology and the law, social intelligence, and normative multi-agent systems to address the ways in which the current information revolution

affects the basic pillars of today's legal and political systems. Also, the Doctoral Consortium attracted additional papers and aimed to help young researchers enter the JURIX community.

Finally, we have the honour to thank the people who have contributed to make JURIX 2017 a success: the colleagues who supported local organisation; Tom van Engers and his Doctoral Consortium committee who worked with doctoral students on their submissions; the reviewers and sub-reviewers who ensured a strict but fair reviewing process; the authors who have submitted papers; the workshop organisers who added auxiliary meetings beyond the central programme of the main conference; and last but not least, the members of the JURIX Steering Committee as well as the current JURIX board who guide JURIX over the year.

Adam Wyner – JURIX 2017 Programme Chair
Giovanni Casini – JURIX 2017 Local Organisation

Conference Organisation

PC Chair
Adam Wyner, University of Aberdeen

Local Chair
Giovanni Casini, University of Luxembourg

Doctoral Consortium Chair
Tom van Engers, University of Amsterdam

Local Organization Committee
Magali Martin, University of Luxembourg
Jérémie Dauphin, University of Luxembourg
Xavier Parent, University of Luxembourg
Livio Robaldo, University of Luxembourg

Program Committee
Michał Araszkiewicz, Jagiellonian University
Kevin Ashley, University of Pittsburgh
Katie Atkinson, University of Liverpool
Trevor Bench-Capon, University of Liverpool
Floris Bex, Utrecht University
Alexander Boer, University of Amsterdam
Karl Branting, The MITRE Corporation
Elena Cabrio, University of Cote d'Azur
Pompeu Casanovas, UAB
Jack Conrad, Thomson Reuters
Tom van Engers, University of Amsterdam
Enrico Francesconi, ITTIG-CNR
Randy Goebel, University of Alberta
Tom Gordon, University of Postdam
Guido Governatori, NICTA
Matthias Grabmair, University of Pittsburgh
Davide Grossi, University of Liverpool
Rinke Hoekstra, Elsevier
John Joergensen, Rutgers University
Yoshinobu Kano, Shizuoka University
Jeroen Keppens, King's College London
Thorne McCarty, Rutgers University
Adeline Nazarenko, University of Paris 13
Katsumi Nitta, Tokyo Institute of Technology
Paulo Novais, University of Minho
Marc van Opijnen, KOOP
Monica Palmirani, CIRSFID

Wim Peters, University of Sheffield
Radim Polčák, Masaryk University
Henry Prakken, Univ of Utrecht and Faculty of Law at Univ of Groningen
Guilin Qi, Southeast University
Alexandre Rademaker, IBM Research Brazil and EMAp/FGV
Livio Robaldo, University of Luxembourg
Anna Ronkainen, University of Helsinki
Antonino Rotolo, University of Bologna
Giovanni Sartor, EUI/CIRSFID
Ken Satoh, National Institute of Informatics and Sokendai
Burkhard Schafer, University of Edinburgh
Fernando Schapachnik, Universidad de Buenos Aires
Erich Schweighofer, University of Vienna
Sarah Sutherland, Canadian Legal Information Institute
Leon van der Torre, University of Luxembourg
Bart Verheij, University of Groningen
Serena Villata, CNRS - Sophia-Antipolis
Radboud Winkels, University of Amsterdam
Peng Xia, Shanghai Bestone Information Technology Co. Ltd
Tomasz Zurek, Maria Curie-Sklodowska University in Lublin

JURIX steering committee

Pompeu Casanovas, Universitat Autonomá de Barcelona
Monica Palmirani, University of Bologna
Erich Schweighofer, University of Vienna
Serena Villata, CNRS

JURIX executive committee

Tom van Engers, University of Amsterdam, president
Bart Verheij, University of Groningen, vice-president/secretary
Floris Bex, Utrecht University, treasurer

Contents

Preface v
Adam Wyner and Giovanni Casini

Conference Organisation vii

Normative Requirements as Linked Data 1
Fabien Gandon, Guido Governatori and Serena Villata

Classifying Legal Norms with Active Machine Learning 11
Bernhard Waltl, Johannes Muhr, Ingo Glaser, Georg Bonczek, Elena Scepankova and Florian Matthes

Cloudy with a Chance of Concepts 21
Suzanne Bardelmeijer, Alexander Boer and Radboud Winkels

Dimensions and Values for Legal CBR 27
Trevor Bench-Capon and Katie Atkinson

Timed Contract Compliance Under Event Timing Uncertainty 33
María-Emilia Cambronero, Luis Llana and Gordon J. Pace

Detecting Agent Mentions in U.S. Court Decisions 39
Jaromír Šavelka and Kevin D. Ashley

Temporalised Belief Revision in the Law 49
Luciano H. Tamargo, Diego C. Martinez, Antonino Rotolo and Guido Governatori

Giving Every Case Its (Legal) Due – The Contribution of Citation Networks and Text Similarity Techniques to Legal Studies of European Union Law 59
Yannis Panagis, Urška Šadl and Fabien Tarissan

Argument Schemes for Discussing Bayesian Modellings of Complex Criminal Cases 69
Henry Prakken

Noise Induced Hearing Loss: An Application of the Angelic Methodology 79
Latifa Al-Abdulkarim, Katie Atkinson, Trevor Bench-Capon, Stuart Whittle, Rob Williams and Catriona Wolfenden

Passing the Brazilian OAB Exam: Data Preparation and Some Experiments 89
Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler and Alexandre Rademaker

Answering Legal Research Questions About Dutch Case Law with Network Analysis and Visualization 95
Dafne van Kuppevelt and Gijs van Dijck

On Annotation of the Textual Contents of Scottish Legal Instruments 101
Adam Wyner, Fraser Gough, Francois Levy, Matt Lynch and Adeline Nazarenko

Balancing with Thresholds 107
Michał Araszkiewicz and Tomasz Zurek

Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links 113
Tommaso Agnoloni, Lorenzo Bacci, Ginevra Peruginelli, Marc van Opijnen, Jos van den Oever, Monica Palmirani, Luca Cervone, Octavian Bujor, Arantxa Arsuaga Lecuona, Alberto Boada García, Luigi Di Caro and Giovanni Siragusa

Scoring Judicial Syllabi in Portuguese 119
Jean-Rémi Bourguet and Melissa Zorzanelli Costa

A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases 125
Isar Nejadgholi, Renaud Bougueng and Samuel Witherspoon

Toward Building a Legal Knowledge-Base of Chinese Judicial Documents for Large-Scale Analytics 135
Amarnath Gupta, Alice Z. Wang, Kai Lin, Haoshen Hong, Haoran Sun, Benjamin L. Liebman, Rachel E. Stern, Subhasis Dasgupta and Margaret E. Roberts

Automated Detection of Unfair Clauses in Online Consumer Contracts 145
Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Yannis Panagis, Giovanni Sartor and Paolo Torroni

A Deep Learning Approach to Contract Element Extraction 155
Ilias Chalkidis and Ion Androutsopoulos

Automatic Detection of Significant Updates in Regulatory Documents 165
Kartik Asooja, Oscar Ó Foghlú, Breiffni Ó Domhnaill, George Marchin and Sean McGrath

A Computational Model of Moral and Legal Responsibility via Simplicity Theory 171
Giovanni Sileno, Antoine Saillenfest and Jean-Louis Dessalles

Toward Linking Heterogenous References in Czech Court Decisions to Content 177
Jakub Harašta and Jaromír Šavelka

Utilizing Vector Space Models for Identifying Legal Factors from Text 183
Mohammad H. Falakmasir and Kevin D. Ashley

Concept Recognition in European and National Law 193
Rohan Nanda, Giovanni Siragusa, Luigi Di Caro, Martin Theobald, Guido Boella, Livio Robaldo and Francesco Costamagna

Subject Index 199

Author Index 201

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-1

Normative Requirements as Linked Data

Fabien GANDON [a,1,2], Guido GOVERNATORI [b] and Serena VILLATA [a]
[a] *Université Côte d'Azur, Inria, CNRS, I3S, France*
[b] *Data61, CSIRO, Australia*

Abstract. In this paper, we propose a proof of concept for the ontological representation of normative requirements as Linked Data on the Web. Starting from the LegalRuleML ontology, we present an extension of this ontology to model normative requirements and rules. Furthermore, we define an operational formalization of the deontic reasoning over these concepts on top of the Semantic Web languages.

Keywords. Linked data, Semantic Web, Deontic rules, Ontology

1. Introduction

The Linked Data principles [3] provide a standard approach to weave a Web of data, linking datasets across the world and virtually in any domain. The semantic Web frameworks additionally provide standard means to publish data (RDF [4]), ontological knowledge (RDFS [5] and OWL [6] schemata), and to query and reason on them (SPARQL [7]). Despite existing approaches to model legal ontological knowledge [9,1,2], little work has been devoted towards the definition of an end-to-end framework to represent, publish and query ontological knowledge from the legal domain using such standards. In this paper, we study how Semantic Web frameworks could apply to the formalization, publication and processing of legal knowledge, and in particular, *normative requirements* and *rules*.

A linked data based deontic representation and reasoning allows us to (a) rely on Web standard to represent, exchange and foster interoperability between deontic rule bases and reasoning systems, (b) rely on existing standards (e.g. SPARQL) and infrastructures (e.g. triple stores) to implement deontic systems, and (c) combine linked data and semantic Web reasoning and formalisms (e.g., OWL) with deontic reasoning to support more inferences.

Our research question is: *Can we represent and reason on the deontic aspects of normative rules with standard Semantic Web languages?* We focus here on two sub-questions: *For which aspects schema-based reasoning (RDFS, OWL) is relevant?* and *Can we operationally formalize other deontic reasoning rules with RDF and SPARQL?*

We first survey the related work to show that current legal vocabularies on the Semantic Web do not provide the expressiveness we need (Section 2). Then we specify and formalize of the ontology we require (Section 3). We describe how normative requirements

[1]Corresponding Author: Fabien Gandon, Inria, Wimmics, 2004 rt des Lucioles, 06902 France; E-mail: fabien.gandon@inria.fr, http://fabien.info.
[2]The authors have received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts".

can be represented as Linked Data (Section 4), and why the states of affairs should be represented as RDF 1.1 named graphs (Section 5). Relying on this modeling, we show that some aspects of deontic reasoning cannot be covered by the OWL formalization whilst they can be captured with SPARQL rules (Section 6). We experiment this approach with a proof of concept (Section 7) before concluding.

2. Related Work

We performed a search[3] on LOV [8], a directory of Semantic Web vocabularies and schemata, to see how legal concepts are covered in published ontologies. Among the retrieved vocabularies, we identified that:

- the General Ontology for Linguistic Description (GOLD) includes a "Deontic Modality" concept [4] but it is essentially defined from a linguistic point of view with the goal to perform natural language analysis.
- the Public Procurement Ontology (PPROC) has the notion of "Contract additional obligations" which is a class limited to describing the additional obligations a contract requires[5].
- the Open Standards for Linking Governments Ontology (OSLO) includes an upper class "permission", but attached to the role of an individual in a society[6].
- the notions of rights, permissions and licenses are mentioned in schemata such as Dublin Core[7], Creative Commons[8] or ODRL[9] but to describe the possible uses of a digital resource and they remain at a descriptive non-formalized level.

Current ontologies are often limited to a specific domain of application and have very shallow coverage of deontic concepts. They are not designed with the goal to support deontic reasoning above Semantic Web frameworks. Their primitives are designed to annotate resources with the goal of documenting or supporting some degree of interoperability, but they are not intended to support Semantic Web based reasoning and processing of the normative requirements and rules. Closer to our goal is the LegalRuleML Meta Model [9] providing primitives for deontic rule and normative requirement representation (Permission, Obligation, Prohibition). We started from this model and extended it with a new ontology focusing on the deontic aspects, integrating notions from an existing abstract formal framework for normative requirements of regulatory compliance [10], and previous on modal defeasible reasoning for deontic logic on the Semantic Web [11]

3. Ontological extension of the LegalRuleML Meta Model

In this section, we first describe the competency questions that motivate our extension of the LegalRuleML ontology, and then we detail the core concepts of our new legal ontology as well as their formalization in OWL.

[3] Keywords include: obligation, prohibition, permission, rights and licences.
[4] http://purl.org/linguistics/gold/DeonticModality
[5] http://linguistics-ontology.org/gold/DeonticModality
[6] http://purl.org/oslo/ns/localgov#Permission
[7] http://dublincore.org/
[8] https://creativecommons.org/ns
[9] http://w3c.github.io/poe/vocab/

3.1. Motivating scenarios and competency questions

Among the many approaches to design an ontology [12], the writing of motivating scenarios is a very usual initial step of specifications to capture problems that are not adequately addressed by existing ontologies [13]. The motivating scenario for us here is to support the annotation, detection and retrieval of normative requirements and rules. We want to support users in information retrieval with the ability to identify and reason on the different types of normative requirements and their statuses. This would be possible through ontology population approaches, but the lack of an existing ontology covering these aspects slows this process, as well as the further development of more advanced applications in legal computer science.

In a second step of ontology specification, a standard way to determine the scope of the ontology is to extract from the scenarios the questions and answers it should be able to support if it becomes part of knowledge-based system. These so-called *competency questions* [13] place demands on the targeted ontology, and they provide expressiveness requirements. The competency questions we target for this ontology are:

- What are the instances of a given requirement and its sub-types, e.g. obligation?
- Is a requirement violated by one or more states of affairs, and if so, which ones?
- Is a given description of rules and states of affairs coherent?
- Which rules, documents and states of affairs are linked to a requirement and how?

3.2. Core primitives

To support the competency questions and relying on definitions from LegalRuleML [9] and deontic reasoning [10,11], we identified a set of core primitives for an ontology capturing the different aspects of normative requirements, and supporting the identification and classification tasks. We called that ontology Normative Requirement Vocabulary (NRV), and made it available and dereferenceable following the Linked Data principles. The namespace is `http://ns.inria.fr/nrv#` with the preferred prefix `nrv` respectively submitted both to LOV [8] and to `http://prefix.cc`.

The top class of the ontology is the `Normative Requirement` which is defined as the set of the requirements implying, creating, or prescribing a norm. Then we have a number of upper classes to capture different features of the requirements:

- `Compensable Requirement`, `Non Compensable Requirement`, `Compensated Requirement` are classes of requirements with different compensation statuses.
- the classes `Violable requirement`, `Non Violable Requirement`, `Violated Requirement` and `Compliant Requirement` characterize the requirements with respect to their relation to a `Compliance` or a `Violation`.
- the other classes follow the same logic, and they distinguish requirements with respect to their perdurance, persistence, co-occurance and preemptiveness.

Using these upper classes, we positioned and extended three primitives from the LegalRuleML Meta Model (i.e., `Prohibition`, `Permission`, `Obligation`), each one inheriting from the appropriate super classes we introduced. For instance, `Permission` inherits from `Non Violable Requirement` and `Non Compensable Requirement`, while `Obligation` inherits from `Violable Requirement` and `Compensable Requirement`. Specializations of these classes are then used to introduce the notions of `Achievement`, `Maintenance` and

`Punctual`. For the complete list of classes and their definitions, we refer the reader to the online documentation available at the namespace URL. These primitives and definitions provide the taxonomic skeleton of our NRV ontology.

3.3. Formalization

In this section, we provide some formalization details (ontological commitment) and their translation into OWL (computational commitment). We will use the TriG syntax [14] for RDF, and the prefixes we use in the rest of this article are:

```
lrmlmm: http://docs.oasis-open.org/legalruleml/ns/v1.0/metamodel#
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
rulemm: http://docs.oasis-open.org/legalruleml/ns/v1.0/rule-metamodel#
xml: http://www.w3.org/XML/1998/namespace
xsd: http://www.w3.org/2001/XMLSchema#
nrv: http://ns.inria.fr/nrv#
nru: http://ns.inria.fr/nrv-inst#
```

We captured the disjointedness expressed in the upper classes representing exclusive characteristics of normative requirements (compensable / non-compensable, violable / non-violable, persistent / non persistent):

```
:NormativeRequirement a rdfs:Class ;
  owl:disjointUnionOf ( :CompensableRequirement :NonCompensableRequirement ) ;
  owl:disjointUnionOf ( :ViolableRequirement :NonViolableRequirement ) ;
  owl:disjointUnionOf ( :PersistentRequirement :NonPersistentRequirement ) .
```

We initially considered the disjointedness of a compliant requirement and a violated requirement, however this disjointedness is not global but local to a state of affairs and therefore it does not translate to a general disjointedness of classes, i.e., a requirement may be violated by a state of affairs but compliant with an other one at the same time. However, this led us to capture this issue as a property disjointedness, since a requirement cannot be violated and be compliant with the same state of affairs at the same time:

```
:hasCompliance a owl:ObjectProperty ; rdfs:label "has for compliance"@en ;
 rdfs:domain :ViolableRequirement ; rdfs:range lrmlmm:Compliance ;
 owl:propertyDisjointWith :hasViolation .
```

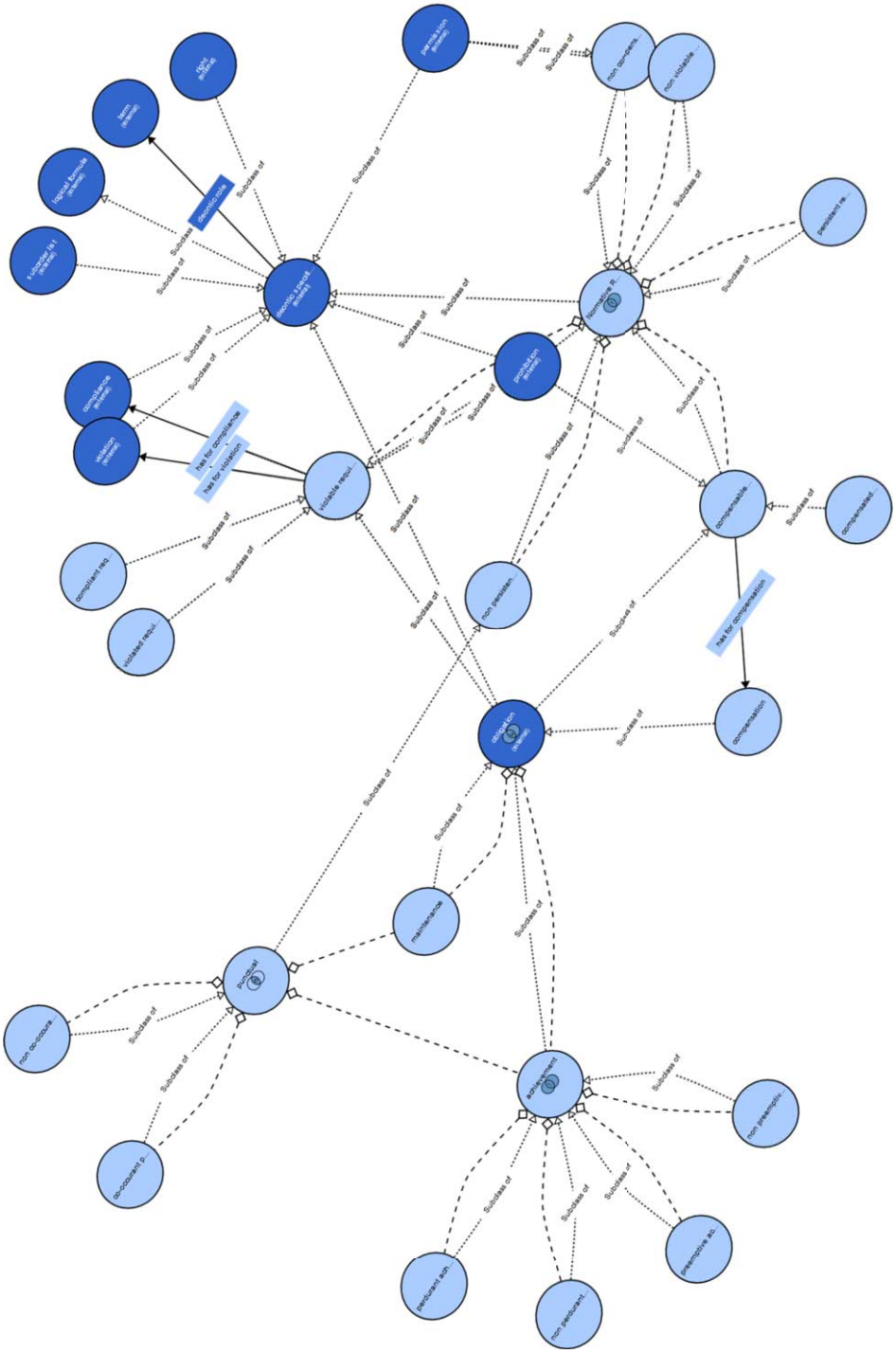
Obligations are an example of non disjoint union between achievements and maintenances, since a punctual requirement is both an achievement and a maintenance:

```
lrmlmm:Obligation a rdfs:Class ;
  rdfs:subClassOf :ViolableRequirement ;
  rdfs:subClassOf :CompensableRequirement ;
  owl:unionOf ( :Achievement :Maintenance )  .

:Achievement a rdfs:Class ; rdfs:label "achievement"@en ;
  owl:disjointUnionOf ( :PreemptiveAchievement :NonPreemptiveAchievement ) ;
  owl:disjointUnionOf ( :PerdurantAchievement :NonPerdurantAchievement ) ;
  rdfs:subClassOf lrmlmm:Obligation .

:Maintenance a rdfs:Class ; rdfs:label "maintenance"@en ;
  rdfs:subClassOf lrmlmm:Obligation .
```

Figure 1. Overview of the NRV ontology and its core primitives, in particular `Prohibition`, `Permission`, `Obligation` and a number of upper classes to capture different features of a `Normative Requirement`.



Violated and compensated requirements could be defined with restrictions on the properties `hasViolation` and `hasCompensation`:

```
:ViolatedRequirement a rdfs:Class ;
    rdfs:subClassOf :ViolableRequirement ;
    owl:equivalentClass  [ a owl:Restriction ;
       owl:onProperty :hasViolation ;
       owl:minCardinality 1 ] .

:CompensatedRequirement a rdfs:Class ;
    rdfs:subClassOf :CompensableRequirement ;
    owl:equivalentClass  [ a owl:Restriction ;
       owl:onProperty :hasCompensation ;
       owl:minCardinality 1 ] .
```

We could now be tempted to define a compliant requirement with the following restrictions:

```
1 :CompliantRequirement a rdfs:Class ; rdfs:label "compliant requirement"@en ;
2   rdfs:subClassOf :ViolableRequirement ;
3   owl:equivalentClass  [ a owl:Restriction ;
4     owl:onProperty :hasCompliance ;
5     owl:minCardinality 1 ] .
6   owl:equivalentClass  [ a owl:Restriction ;
7     owl:onProperty :hasViolation ;
8     owl:maxCardinality 0 ] .
```

However we removed the second part (lines 6-8) of the restriction since it re-introduces a disjunction between the compliant and violated requirement classes. The notions of compliance and violation are not generally disjoint but only disjoint locally to a state of affair, i.e., a normative requirement can be violated and compliant at the same time but with respect to different states of affairs. However, OWL definitions cannot rely on RDF 1.1 named graphs, which we will use for representing states of affairs. Therefore we will need another mechanism to capture this kind of constraints.

Because we used disjoint unions, the ontology is in OWL DL, i.e., $\mathcal{SHOIN}^{(\mathcal{D})}$, more precisely in the $\mathcal{AL}(\mathcal{U})\mathcal{C}(\mathcal{H})\mathcal{RN}$ family, i.e., $\mathcal{AL}$ attributive language, ($\mathcal{U}$ concept union), $\mathcal{C}$ complex concept negation, ($\mathcal{H}$ role hierarchy), $\mathcal{R}$ limited complex role inclusion axioms, reflexivity, irreflexivity, role disjointedness, and $\mathcal{N}$ cardinality restrictions.

We decided to declare the signature of properties (e.g., `hasViolation`, `hasCompensation`) at the ability level (e.g., violable requirement, compensable requirement), and not at the effective status level (e.g., violated requirement, compensated requirement) because each status will be local to a state of affairs. Therefore, in the end, we avoided too strong restrictions and signatures. If we remove cardinality restrictions, unions and disjointedness, the ontology becomes compatible with OWL EL and OWL RL which could be interesting for implementations relying on rule-based systems, especially when we consider the extensions proposed in the following sections.

4. Requirements as Linked Data

Using the LegalRuleML Meta Model and the NRV ontology we can now start to represent normative requirements as Linked Data. Let us introduce two examples. The first one is a rule stating that according to Australian law one cannot drive over 90km/h:

```
<http://gov.au/driving-rule> a lrmlmm:Source ;
   rdfs:label "driving rules in Australia"@en .
nru:LSS1 a lrmlmm:Sources ;
   lrmlmm:hasLegalSource <http://gov.au/driving-rule> .
nru:LRD1 a lrmlmm:LegalRuleMLDocument ;
   lrmlmm:hasLegalSources nru:LSS1 ;
   lrmlmm:hasAlternatives [ lrmlmm:fromLegalSources nru:LSS1 ;
                            lrmlmm:hasAlternative nru:PS1 ] ;
   lrmlmm:hasStatements nru:SS1 .
nru:SS1 a lrmlmm:Statements ;
  lrmlmm:hasStatement nru:PS1 .
nru:PS1 a lrmlmm:PrescriptiveStatement, lrmlmm:Prohibition ;
  rdfs:label "can't drive over 90km"@en .
```

The second example is a rule stating that employees of CSIRO must wear their badges:

```
<http://csiro.au/security-rule> a lrmlmm:Source ;
  rdfs:label "security rules in CSIRO"@en .
nru:LSS2 a lrmlmm:Sources ;
  lrmlmm:hasLegalSource <http://csiro.au/security-rule>  .
nru:LRD2 a lrmlmm:LegalRuleMLDocument ;
  lrmlmm:hasLegalSources nru:LSS2 ;
  lrmlmm:hasAlternatives [ lrmlmm:fromLegalSources nru:LSS2 ;
                           lrmlmm:hasAlternative nru:PS2 ] ;
  lrmlmm:hasStatements nru:SS2 .
nru:SS2 a lrmlmm:Statements ;
  lrmlmm:hasStatement nru:PS2 .
nru:PS2 a lrmlmm:PrescriptiveStatement, lrmlmm:Obligation ;
  rdfs:label "you must wear your badge inside CSIRO facilities"@en .
```

5. State of affairs as named graphs.

The ability to define contexts and group assertions was one of the main motivations for having named graphs in RDF 1.1 [15]. The notion of state of affairs at the core of deontic reasoning is naturally captured by named graphs where all the statements of each state of affairs are encapsulated as RDF triples in a named graph, identifying that precise state of affairs. We provide here four examples of states of affairs respecting (2 and 3) or breaking (1 and 4) the rules of the normative statements described above. The core idea is to represent each state of affairs as a named graph typed as a factual statement of LegalRuleML.

```
:StateOfAffairs1 a lrmlmm:FactualStatement .
GRAPH :StateOfAffairs1 { rdfs:label "Tom" ;
  :Tom :activity [ a :Driving ;
                   :speed "100"^^xsd:integer ;
                   rdfs:label "driving at 100km/h"@en ] . }
:StateOfAffairs2 a lrmlmm:FactualStatement .
GRAPH :StateOfAffairs2 {
  :Jim :activity [ a :Driving ; rdfs:label "Jim" ;
                   :speed "90"^^xsd:integer ;
                   rdfs:label "driving at 90km/h"@en ] . }
:StateOfAffairs3 a lrmlmm:FactualStatement .
GRAPH :StateOfAffairs3 { rdfs:label "Jane" ;
 :Jane :location [ rdf:value :CSIRO ;
                   :start "2017-07-18T09:30:10+09:00"^^xsd:date ;
```

```
                    :end   "2017-07-18T17:00:10+09:00"^^xsd:date ] ;
        :badge   [ rdf:value :CSIRO ;
                    :start "2017-07-18T09:30:10+09:00"^^xsd:date ;
                    :end   "2017-07-18T17:00:10+09:00"^^xsd:date ] . }
:StateOfAffairs4 a lrmlmm:FactualStatement .
GRAPH :StateOfAffairs4 { rdfs:label "Steve" ;
 :Steve :location [ rdf:value :CSIRO ;
                    :start "2017-07-18T09:30:10+09:00"^^xsd:date ;
                    :end   "2017-07-18T17:00:10+09:00"^^xsd:date ] ;
        :badge    [ rdf:value :CSIRO ;
                    :start "2017-07-18T10:30:10+09:00"^^xsd:date ;
                    :end   "2017-07-18T17:00:10+09:00"^^xsd:date ] . }
```

6. Deontic reasoning as SPARQL rules

Since the notion of named graph that appeared with RDF 1.1 (2014, [4]) is absent from OWL 2 (2012, [6]) and its constructors, we need to implement the reasoning on states of affairs by other means. The SPARQL language is both a standard and a language able to manipulate named graphs so we propose to use SPARQL rules. In this section, we explore the coupling of OWL reasoning with SPARQL rules to formalize and implement some deontic reasoning. Description Logics (DL) support reasoning on the description of concepts and properties of a domain (terminological knowledge or T-Box) and of their instances (assertional knowledge or A-box). They are the basis of the Web Ontology Language (OWL). The classical inferences supported by DL are instance checking, relation checking, subsumption checking, and consistency checking [16]. While these inferences are useful to reason about deontic knowledge (e.g., a compensable requirement must also be a violable requirement), they do not cover all the inferences we want to support here in particular deontic rules (e.g., a requirement is violated by a state of affairs if, during a specific period of time, a given constraint does not hold). These rules rely on complex pattern matching including, for instance, temporal interval comparison that go beyond OWL expressiveness. As a proof of concept, the following rules check the violation or compliance of the statements made by the previous states of affairs. The core idea is to add to each named graph of each state of affairs the deontic conclusions of the legal rules relevant to it. By relevant we mean here that the state of affairs describes a situation that falls under the application conditions of that legal rule. The following rules update compliance and violation for the driving speed requirement:

```
DELETE {  graph ?g { nru:PS1 nrv:hasCompliance ?g } }
INSERT {  graph ?g { nru:PS1 a nrv:ViolatedRequirement ;
                     nrv:hasViolation ?g } }
WHERE  {  graph ?g {  ?a a :Driving ; :speed ?s .  }
          FILTER (?s>90) }  ;
DELETE { graph ?g { nru:PS1 a nrv:ViolatedRequirement ;
                    nrv:hasViolation ?g } }
INSERT { graph ?g { nru:PS1 nrv:hasCompliance ?g } }
WHERE  { graph ?g {  ?a a :Driving ; :speed ?s . }
         FILTER (?s<=90) }
```

The following rules update compliance and violation for the CSIRO badge requirement:

```
INSERT { graph ?g { nru:PS2 a nrv:ViolatedRequirement ; nrv:hasViolation ?g }}
WHERE  { graph ?g { ?x :location [ rdf:value ?o ; :start ?ls ; :end ?le ]
```

```
        optional { ?x :badge    [ rdf:value ?o ; :start ?bs ; :end ?be ] .
        FILTER (?bs<=?ls && ?be>=?le) } }
       FILTER (  ( ! bound (?bs))   ) }     ;
INSERT { graph ?g { nru:PS2 nrv:hasCompliance ?g } }
WHERE  { graph ?g { ?x :location [ rdf:value ?o ; :start ?ls ; :end ?le ]
                    ?x :badge    [ rdf:value ?o ; :start ?bs ; :end ?be ] .  }
         FILTER (?bs<=?ls && ?be>=?le) }
```

The following rules update compliance for the state of affairs after violations were checked:

```
INSERT {  graph ?g {?n a nrv:CompliantRequirement }  }
WHERE  {  ?g  a lrmlmm:FactualStatement .
          ?n a nrv:ViolableRequirement .
          graph ?g {  ?n  nrv:hasCompliance ?g }
          minus { graph ?g {  ?n  nrv:hasViolation ?g } } }   ;
DELETE {  graph ?g {?n a nrv:CompliantRequirement }  }
WHERE  {  ?g  a lrmlmm:FactualStatement .
          ?n a nrv:ViolableRequirement .
          graph ?g {  ?n  nrv:hasViolation ?g } }
```

7. Proof of concept and experimentation

To validate and experiment with the ontology, the Linked Data and the rules, we used two established tools:

- the latest version of the Protégé platform [17] and the reasoners it includes were used to check the NRV OWL ontology which was found coherent and consistent.
- the latest version of CORESE [18] was used to load the LegalRuleML and NRV ontologies, the Linked Data about the rules and the states of affairs, and the SPARQL rules to draw the conclusions as shown in Figure 2 for the two first states of affairs concerning speed limitation.

Figure 2. Extract of the quadruples (N-Quads) produced by CORESE after all the reasoning on the two first states of affairs concerning speed limitation showing one violated state (white background) and one compliant one (blue background). The columns indicate the named graph of the state of affairs (`?g`), the subjects (`?lx`), the predicates (`?lp`), and the objects (`?lv`) of the triples in this named graph.

?g	?lx	?lp	?lv
http://ns.inria.fr/nrv-inst#StateOfAffairs1	Tom	http://ns.inria.fr/nrv-inst#activity	driving at 100km/h
http://ns.inria.fr/nrv-inst#StateOfAffairs1	Tom	http://www.w3.org/2000/01/rdf-schema#label	Tom
http://ns.inria.fr/nrv-inst#StateOfAffairs1	can't drive over 90km	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	violated requirement
http://ns.inria.fr/nrv-inst#StateOfAffairs1	can't drive over 90km	has for violation	http://ns.inria.fr/nrv-inst#StateOfAffairs1
http://ns.inria.fr/nrv-inst#StateOfAffairs1	driving at 100km/h	http://ns.inria.fr/nrv-inst#speed	100
http://ns.inria.fr/nrv-inst#StateOfAffairs1	driving at 100km/h	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://ns.inria.fr/nrv-inst#Driving
http://ns.inria.fr/nrv-inst#StateOfAffairs1	driving at 100km/h	http://www.w3.org/2000/01/rdf-schema#label	"driving at 100km/h"@en
http://ns.inria.fr/nrv-inst#StateOfAffairs2	Jim	http://ns.inria.fr/nrv-inst#activity	driving at 90km/h
http://ns.inria.fr/nrv-inst#StateOfAffairs2	Jim	http://www.w3.org/2000/01/rdf-schema#label	Jim
http://ns.inria.fr/nrv-inst#StateOfAffairs2	can't drive over 90km	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	compliant requirement
http://ns.inria.fr/nrv-inst#StateOfAffairs2	can't drive over 90km	has for compliance	http://ns.inria.fr/nrv-inst#StateOfAffairs2
http://ns.inria.fr/nrv-inst#StateOfAffairs2	driving at 90km/h	http://ns.inria.fr/nrv-inst#speed	90
http://ns.inria.fr/nrv-inst#StateOfAffairs2	driving at 90km/h	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://ns.inria.fr/nrv-inst#Driving
http://ns.inria.fr/nrv-inst#StateOfAffairs2	driving at 90km/h	http://www.w3.org/2000/01/rdf-schema#label	"driving at 90km/h"@en

8. Conclusions

In this paper, we addressed the issue that current vocabularies on the Semantic Web do not provide the expressiveness required to support deontic reasoning on normative

requirements and rules. As a contribution, we specified and formalized an ontology extending LegalRuleML, and we showed how it can be used to represent normative requirements as Linked Data with states of affairs represented as RDF 1.1 named graphs. Relying on this modeling, we proposed an approach based on SPARQL rules to cover some of the deontic aspects outside the expressiveness of OWL 2, and we experiment this approach with a proof of concept based on two established tools of the Semantic Web community. Future work includes extensive population and testing of the ontology on larger datasets and cases. In particular, we intend to go beyond the proof of concept by evaluating this end-to-end approach based on the Semantic Web languages on a business process compliance checking scenario [10]. Extensions of this work also include the possibility to represent differentiated classes of validity that would correspond to the actual structure of our legal system and non-binary modes that would be fit to process proportionality of legal principles. The introduction of a complete rule-based system is part of our future directions as well.

References

[1] Hoekstra, R.; Breuker, J.; Bello, M. D.; and Boer, A. 2007. The LKIF core ontology of basic legal concepts. In *Proc. of the Workshop on Legal Ontologies and AI Techniques*.

[2] Sartor, G.; Casanovas, P.; Biasiotti, M.; and Fernndez-Barrera, M. 2013. *Approaches to Legal Ontologies: Theories, Domains, Methodologies*. Springer.

[3] T. Heath, C. Bizer, *Linked data: Evolving the web into a global data space*, Synthesis lectures on the semantic web: theory and technology 1.1 (2011): 1-136.

[4] World Wide Web Consortium, *RDF 1.1 concepts and abstract syntax*, W3C Recommendation, 25 February 2014

[5] World Wide Web Consortium, *RDF Schema 1.1*, W3C Recommendation, 25 February 2014

[6] World Wide Web Consortium, *OWL 2 Web Ontology Language*, Document Overview (2nd Edition), W3C Recommendation 11 December 2012

[7] World Wide Web Consortium, *SPARQL Query Language for RDF*, W3C Recommendation 15 January 2008

[8] Vandenbussche, Pierre-Yves, et al. "Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web." Semantic Web 8.3 (2017): 437-452.

[9] Athan, Tara, et al., OASIS LegalRuleML, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ACM, 2013

[10] M. Hashmi, G. Governatori, M.T. Wynn, Normative requirements for regulatory compliance: An abstract formal framework. *Omnes Information Systems Frontiers* 18(3), pp. 429-455, 2016

[11] K. Efstratios, N. Bassiliades, G. Governatori, G. Antoniou, A Modal Defeasible Reasoner of Deontic Logic for the Semantic Web, *International Journal on Semantic Web and Information Systems*, (IJSWIS) 7 (2011): 1, doi:10.4018/jswis.2011010102

[12] F. Gandon, *Ontology Engineering: a Survey and a Return on Experience*, RR-4396, INRIA. 2002.

[13] M. Uschold, M. Gruninger, Ontologies: Principles, methods and applications, *The knowledge engineering review*, 11(2), 93–136, 1996, Cambridge University Press.

[14] World Wide Web Consortium, *RDF 1.1 TriG*, W3C Recommendation, 25 February 2014

[15] F. Gandon, O. Corby, Name That Graph or the need to provide a model and syntax extension to specify the provenance of RDF graphs., *W3C Workshop — RDF Next Steps*, Jun 2010, Palo Alto, United States.

[16] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider, *The Description Logic Handbook: Theory, Implementation, Applications*. CUP, 2003.

[17] N.F. Noy, M. Sintek, S. Decker, M. Crubézy, R. W. Fergerson, and M. A. Musen, Creating semantic web contents with protege-2000, *IEEE Intelligent Systems*, 16.2 (2001): 60-71.

[18] O. Corby, R. Dieng-Kuntz and C. Faron-Zucker, Querying the semantic web with corese search engine, *In Proc. of the 16th European Conf. on Artificial Intelligence*, pp. 705-709 (2004), IOS Press.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-11

Classifying Legal Norms with Active Machine Learning

Bernhard WALTL [a], Johannes MUHR [a], Ingo GLASER [a], Georg BONCZEK [a], Elena SCEPANKOVA [a], and Florian MATTHES [a]

[a] *Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Germany*

Abstract. This paper describes an extended machine learning approach to classify legal norms in German statutory texts. We implemented an active machine learning (AML) framework based on open-source software. Within the paper we discuss different query strategies to optimize the selection of instances during the learning phase to decrease the required training data.

The approach was evaluated within the domain of tenancy law. Thereby, we manually labeled the 532 sentences into eight different functional types and achieved an average F1 score of 0.74. Comparing three different classifiers and four query strategies the classification performance F1 varies from 0.60 to 0.93. We could show that in norm classification tasks AML is more efficient than conventional supervised machine learning approaches.

Keywords. norm classification, active machine learning, text mining

1. Introduction

More and more textual data that is relevant for the legal domain is digitally available. Algorithms and technological infrastructure for text mining and natural language processing are becoming more powerful in terms of their accuracy and performance. The use cases and tools for text mining in the legal field that are relevant for legal experts or practitioners, e.g., scientists, lawyers, judges, courts, etc., are manifold. A recent overview was published by Ashley in 2017 [1].

From an algorithmical point of view two major approaches exist to structure textual data: rule-based (knowledge-based) approaches and machine learning (ML) (statistical). Both approaches are attractive and have their specific advantages and disadvantages. Nowadays, rule-based approaches are still more common in practice, although science focuses much more on ML (see [2]). Many different notions of ML exist that can be applied to classify, categorize, predict, or cluster textual data. Thereby, active machine learning (AML) seems to be highly attractive, since it decreases the effort of training by providing mechanisms to train ML classifiers more efficiently [3].

This paper describes the combination of rule-based text mining with AML, a specific form of semi-supervised ML, for the classification of legal norms. The

reminder of the paper is structured as follows: Section 2 provides a short overview of the related work, Section 3 describes the architecture of the AML approach, the dataset and used labels are discussed in Section 4, finally the approach and its performance is evaluated in Section 5.

2. Related Approaches in Norm and Sentence Classification

Maat and Winkels performed this task for Dutch legislative text [4,5]. Thereby, they achieved a remarkable accuracy of more than 90% by classifying 13 different classes using a Support Vector Machine (SVM). They also performed the classification using a context free grammar, i.e., rule-based approach, for the classification (see [5]).

Wyner et al. extracted rules from regulations using JAPE grammar and the GATE framework [6]. They have developed a methodology for the extraction of deontic rules using linguistic rules. The quality of the results is varying, but promising: several categories have been extracted with high precision and recall.

The research group of Ashley, Grabmair and Savelka [7,8] extracted of semantic information from legal documents, e.g., statutory texts and cases. Thereby, they used an Apache UIMA type system to extract legal concepts from vaccine injury decisions (see [7]). Beside these rule-based approaches they investigated the potentials of interactive ML in classifying relevancy during an analysis task of statutory texts [8]. They were able to show that this can lead to major improvements during classification tasks.

To the best of our knowledge no attempt to classify norms and sentences for statutory texts in Germany using an active or supervised machine learning approach has ever been made before.

3. Active Machine Learning to Classify Legal Norms

3.1. Knowledge Engineering with Rule-based Approaches

Especially for rule-based approaches, linguistic variation as well as vocabulary variety constitute challenges. This holds within a professional language as well as in technical languages. Variations of pronunciation, vocabulary, and inflections steadily occur. Current research is still facing the so-called paraphrasing issue. Two different people phrase the same message by different wording [9]. A knowledge engineer must pay attention to these facts in order to define proper rules. Although, rule-based approaches are not very popular at today's scientific conferences, they are still pre-dominant in practice [2].

3.2. Active Machine Learning

AML is an adapted form of semi-supervised machine learning, in which the training is done in so-called rounds. Within each round a pre-defined amount of instances are manually labelled. The instances are not randomly selected but determined by a mathematical founded query strategy. The process starts by utilizing

random queried instances (seed set) to initially train a classifier model (1). This trained model is used to predict the labels of the unlabeled instances (2). Based on a query strategy, the unlabeled norms are selected by the classifier to distinguish more efficiently between the types (3). Thereby, query strategies are algorithms using the output probabilities/scores of the classifier to calculate an informativeness measure such as the entropy. These instances (e.g. instances having the highest entropy) are presented to a person to be labeled and added to the training set consisting of the random queried instances. The other (not labeled) instances remain in the unlabeled dataset (4). This process is repeated until some kind of stopping criterion (e.g., confidence threshold, maximum number of rounds) is met [3].

We implement our approach with Apache Spark, which is a fast, fault-tolerant and general-purpose open-source cluster computing framework for large-scale data processing. Spark provides an ecosystem consisting of several components managing the basic functionality (e.g., memory management, task scheduling). Apache Spark offers a ML library called MLlib[1] consisting of a variety of efficient and scalable implementations of common ML settings to conduct (semi-)supervised and unsupervised ML. Additionally, MLlib provides so-called ML Pipelines that facilitate the execution of typical ML classification tasks, i.e., preprocessing, feature extraction, and classification.

AML is an iterative and interactive extension of conventional semi-supervised ML. The key hypothesis of AML is that if the learning algorithm can select the data from which it learns, it will perform better with a smaller training set resulting in a more efficient learning.

3.3. Best-of-breed: Combining Rule-based and Active Machine Learning

The discussed approaches can be combined to tackle two challenges. Firstly the generation of labeled datasets that can be used for supervised machine learning techniques in text analysis and secondly the classification of textual data.

As described in Section 3.1, information extraction based on explicitly formulated rules is an effective way of directly integrating the expertise of domain professionals into the process of knowledge engineering. However, generally rules fail to fully capture the broad linguistic variety encountered in natural language.

The combination of (active) ML and rule-based approaches seems suitable to address the aforementioned challenges assuming that rule-based information extraction suffers from low recall but high precision (assuming the rules are written correctly) and (supervised) ML needs large amount of training data for correct inference. Figure 1 shows the structure of the integration of these two approaches, implemented in different software components. The entities extracted with rules bootstrap the active machine learning part, where the domain expert monitors and supports the learning process by providing input for the ML component (see Section 4.4).

[1]https://spark.apache.org/mllib/, last access on 08/24/2017

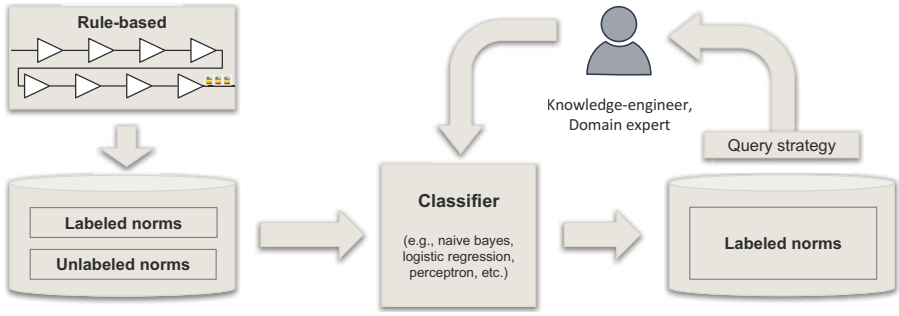


Figure 1. Combining rule-based and AML based approaches for classification of legal norms.

4. Norm Classification with Active Machine Learning

4.1. Objective

The classification of norms is, due to several reasons, attractive for the field of legal informatics. First of all, it allows a more elaborate differentiation of a norm's meaning and thus supports subsequent norm interpretation and formalization. Secondly, it is beneficial for the search and exploration tasks in legal information databases and consequently supports the efficiency of searching of and within legal documents. And finally, it helps determining references and dependencies between and within legal norms.

4.2. Types and Classes

Classification of legal norms can be addressed from different perspectives, e.g., from a philosophical, a legal theoretical or, a constructive one. To achieve the aforementioned tasks—a deeper understanding of interactions between legal norms—, we chose a classification regarding functional types. The taxonomy as well as the gold standards was developed on German statutory legal norms by two legal experts.

In a functional norm classification system, legal norms can be divided into 4 types of statements: normative, auxiliary, legal-technical, legal-mechanism. Our taxonomy comprises normative statements into the following categories: statutory duties, statutory rights, shall-to-do rules and (positive/negative) statutory consequence rules. The taxonomy is shown in Table 1.

The category of statutory duties further comprises the subcategories of order and prohibition, the category of statutory rights is composed of the subcategories of permission and release. The type of auxiliary statement norms can be divided into statements about terms and statements about norms. The first category can be subdivided into explanatory, extending and limiting statements, in which the explanatory statements include the subcategories of definition and precision statements. The category "statement about norms" is subdivided into modifications, legal validity, scope and area of application categories. Where the norms are dominated by their legal-technical or legal-mechanism nature, we identified the categories of reference and continuation in the first section and the categories

Normative statements	Statutory duties		Order
			Prohibition
	Statutory rights		Permission
			Release
	Shall-to-do rules		Shall-to-do rules
	Legal consequences		Legal consequences pos.
			Legal consequences neg.
Auxiliary statements	Statements about terms	Explanatory statements	Definition
			Precision
		Extension and limitations	
	Statement about norms	Legal validity	Legal validity and non-validity
		Scope of application	Temporal
			Personal
			Factual
		Area of application	Extension
			Limitation
			Definition
		Modifications	
Legal-technical statements			Reference
			Continuation
Legal-mechanism statements			Procedure
			Objection

Table 1. Functional type classification of statutory legal norms for Germany's legislative texts.

of procedure and objection in the second section. Table 1 shows 22 types are identified, with considerable differences in their support within the tenancy law.

4.3. Data

In order to prepare a suitable dataset for the norm classification experiment, a legal expert assigned a type to every sentence of the tenancy law section in the German civil code (§535 - §595) published on March 1st, 2017. The result was 532 labeled sentences using 16 different labels. As 16 of the 22 labels had a support less than 1,2%, they were removed from the dataset used. The 504 remaining sentences used for this classification task were composed of the eight classes illustrated in Table 2.

From this dataset, 126 sentences (25%) were randomly added to the test set. The remaining 378 sentences (75%) were used for iterative training. It was ensured that enough instances of each class were in both datasets. We used tokens and their POS tags as features to represent norm instances.

4.4. Experiment and Query Strategies

In this experiment, nine combinations using AML query strategies (see Tables 3 and 4) as well as three combinations using conventional supervised learning (CSL)

Type (German)	Type (English)	Occurrences	Support
Recht	statutory rights	126	25,00%
Pflicht	statutory duties	109	21,63%
Einwendung	objection	92	18,25%
Rechtsfolge	legal consequence	50	9,92%
Verfahren	procedure	49	9,72%
Verweisung	reference	46	9,13%
Fortführungsnorm	continuation	19	3,77%
Definition	definition	13	2,58%

Table 2. Types and statistics of used and manually labeled dataset.

were conducted for each classifier. In CSL, instances are queried randomly without applying any query strategy. These query strategies refer either to uncertainty sampling (US) or to the more elaborated query by committee (QBC) methods. While the former uses only one classifier model, the latter creates a committee of classifiers with the intention to cover a larger area of the version space. To create the classifier committee, the composition of the training data was adapted for each committee. Except for the QBC Vote Entropy strategy, all strategies take advantage of the output probabilities.

Query Strategy	Method	Description
Uncertainty Sampling (US)	Entropy	Selection based on the avg. information content (Shannon entropy) of an instance.
	Margin Sampling (MS)	Selection based on the output margin of the predicted outcomes with the highest prob.
Query by Comittee (QBC)	QBC Vote Entropy (VE)	Selection based on a committee of different QS methods (ensemble with majority vote).
	QBC Soft VE	Selection based on a committee of different QS methods (ensemble with majority vote, including probabilities).

Table 3. Query strategies for active machine learning.

As the MLP does not produce any output score, only the QBC vote entropy approach could be used with this classifier. Each of these twelve combinations was executed five times and averaged to obtain a significant and comparable result.

In the first round, instances (seed set) were randomly queried from the unlabeled training set, labeled and used for learning in the first round. In the subsequent rounds, again either the five most informative instances in the case of a

Abbr.	Classifier	Query Strategy
NB	Multinomial Naive Bayes	Entropy, MS, QBC VE, QBC Soft VE
LR	Logistic Regression	Entropy, MS, QBC VE, QBC Soft VE
MLP	Multi-layer Perceptron	QBC VE

Table 4. Combination of the applied evaluation settings.

query strategy were used; or five random instances were removed from the unlabeled training set, labeled and added to the labeled training set. For both classifiers used (NB and LR), five-fold cross validation was applied to ensure that these predictions were made with the best model found.

After each round, the resulting pipeline model was applied to the test data to evaluate the performance of the current model. This process was repeated until all instances of the training set were labeled (72 learning rounds in total).

4.5. Parametrization of Classifiers

The classifiers NB has been used with standard parametrization of MLLib. Due to performance reasons, the number of iterations for LR was decreased from 100 (default) to 10. The MLP had four layers, whereas the number of nodes of the two intermediate layers was 20 and 10, respectively. The size of the input layer was 2^{13} and the size of the output layer eight (i.e., number of types). The size of the seed set was 18 instances for each of iteration of norm classification.

5. Evaluation

The objective of this experiment was to evaluate (1) the potential of AML compared to CSL and (2) the quality of legal norm classification using ML/AML.

To achieve this, the model was evaluated with an independent test set after each round. To compare the performance of the AML approach, we used standard evaluation metrics[2]: *precision*, *recall*, F_1 and *accuracy*. Additionally, learning curves were utilized to monitor and visualize the learning progress.

None of the four used query strategies had shown to be significantly predominant compared to the others. Thus, the *average accuracy* combines the result of all query strategies used for the classifiers NB and LR, respectively, is visualized in Figure 2. It shows the performance of classifiers applying AML techniques opposed to CSL methods querying random instances.

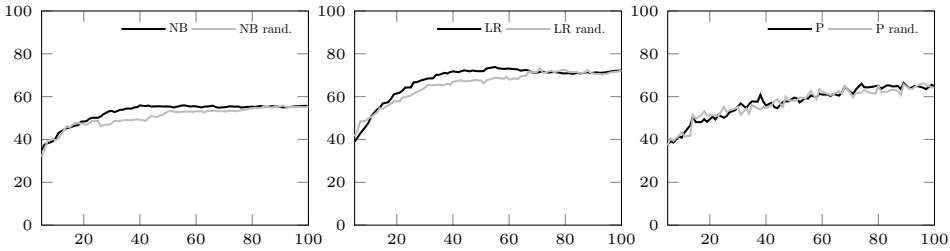


Figure 2. Average accuracy of classifiers vs. random learning (NB=Naive Bayes, LR=Logistic Regression, P=Perceptron.). Y-axis is accuracy in %, X-axis is labeled instances in %.

It becomes evident that AML is clearly superior to CSL when using NB and LR. The use of AML increased not only the speed of learning, but also resulted

[2]Note: no binary classification

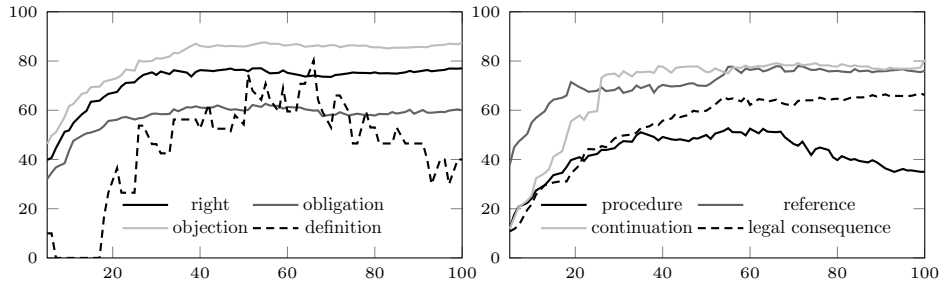


Figure 3. Average F_1 per class (active LR). Y-Axis is F_1 in %, X-axis is labeled instances in %.

in a higher maximum accuracy obtained during the classification process. In both experiments, the average accuracy was after a short "discovery phase" up to 5%-10% higher when having labeled 20%-70% of the instances compared to the random approach. Additionally, the accuracy obtained was higher all instances. Increasing the number of AML rounds, the chance of overfitting is increasing as well, so that after a certain number of labeled instances (70%-95%) both approaches align to the same final accuracy.

When analyzing the results of the individual learning rounds of a specific combination, the importance of having a "high quality seed" set becomes clear. As

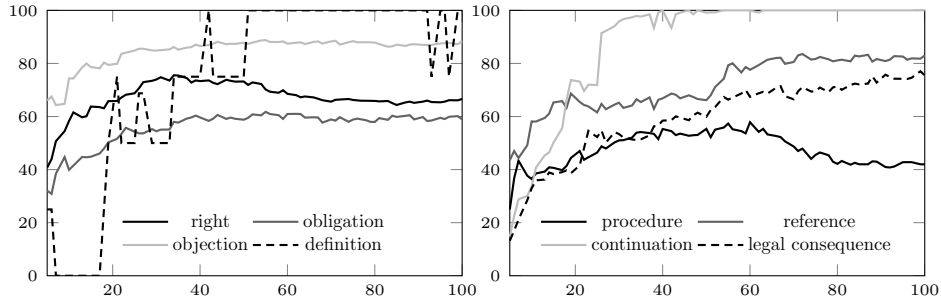


Figure 4. Average precision per type using logistic regression classification. Y-Axis is precision in %, X-axis is labeled instances in %.

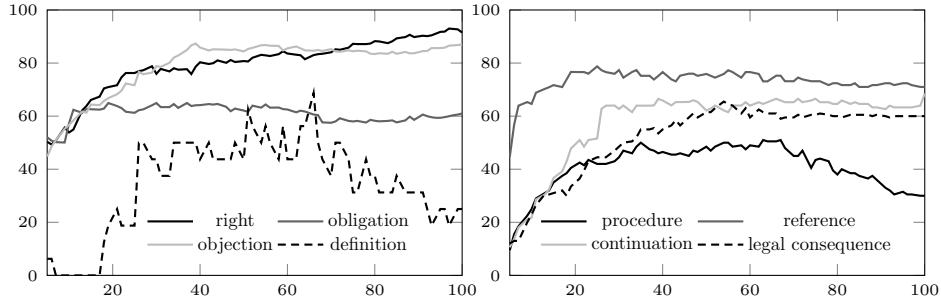


Figure 5. Average recall per type using logistic regression classification. Y-Axis is recall in %, X-axis is labeled instances in %.

the seed set in this study was created randomly for each experiment, the learning differs especially in the initial phases. Only after a discovery of the version space (discovery phase), AML was significantly superior to CSL. An improved coverage of the version space resulted in an almost 20% higher accuracy having labeled only 17% of the instances. Further, a maximum accuracy of almost 80% could be achieved having labeled only 35% of the instances (see Figure 3). An increase of more than 6% compared to CSL using 65% less instances.

To analyze the recognition of individual classes, *consolidated evaluation measures* (averaging the results of all four query strategies) obtained by the LR, the best classifier, are used. Figure 3, 4 and 5 show the consolidated curves.

Thereby, the different (final) results of the individual labels are very noticeable. While norms belonging to the type *objection* are well recognized, soon having an F_1 of almost 90%, towards the end, norms referring to the type *definition* or *procedure* cannot be classified easily by a classifier. The reason for the low end-value for the type *definition* might be their low support - with less than 3% - resulting in an only very small training set. Despite the fact that the training set for the type *continuation* contains only two more instances this type has an F_1 value of more than 80%. Thus, the classifier might also have problems to distinguish a *definition* from other types or the kind of *definitions* in the training set is linguistically varying from the one of the test set (different sub-type).

However, considering the intermediate results, the types *continuation* and *definition* that have only a very low support in the dataset, have both a very high precision and also a good recall temporarily. Hence, the reason for the worsening results is more likely caused by the overfitting of the classifier. This can be confirmed by the results attained by the type *procedure* that achieves much better results during the classification process. Nevertheless, this type shows the worst results having both a low precision and recall. Although the number of training instances is high for the type *obligations*, the classifier has problems recognizing them in the test set. The norm type *right* had the highest recall towards the end, but a rather low precision (see Figures 4 and 5).

6. Outlook and Future Work

This work is an additional step towards supervised machine learning with the objective to decrease the effort of labeling. Based on the results of this study, we see several next steps that can be addressed: i) Deep investigation of the reasons why the F1 measure for different norm types differ so heavily? ii) How do comparably low support of norm types (e.g., definitions) effect the classifier and how can negative impacts be avoided? iii) Does the full-stack integration of AML and rule-based approaches lead to even better performance and faster learning?

Beside these technical questions it would be interesting to adapt and apply this method to statutory (or judicial) texts of foreign languages, e.g. english. This could support current ongoing research projects, e.g. [8,10].

7. Summary

This paper describes active machine learning to classify legal norms in German statutory texts. Thereby, the classifier is trained in multiple rounds using a mathematical function, i.e. query strategy, which selects the most informative instances. This leads to an efficient learning for the classifier an minimizes the required training data.

Based on a functional type classification of legal norms we evaluated the approach in the field of German tenancy law. We compared three classifiers and four different query strategies in 72 learning rounds. For certain norm types, e.g., objections, rights, and obligations, a high detection accuracy of about 0.90 was achieved.

We consider this as a fruitful research direction to decrease the efforts required in supervised machine learning approaches for legal text classification.

Acknowledgment

This research was partially funded by an R&D grant on Contract Analysis by SINC GmbH, Wiesbaden, Germany.

References

[1] K. D. Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge: Cambridge University Press, 2017.

[2] L. Chiticariu, Y. Li, and F. R. Reiss, "Rule-based information extraction is dead! long live rule-based information extraction systems!" in *EMNLP*, no. October, 2013, pp. 827–832.

[3] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, no. 55-66, p. 11, 2010.

[4] E. de Maat, K. Krabben, and R. Winkels, "Machine Learning versus Knowledge Based Classification of Legal Texts," in *JURIX*, 2010, pp. 87–96.

[5] E. de Maat and R. Winkels, "Automated Classification of Norms in Sources of Law," in *Semantic processing of legal texts*, E. Francesconi, Ed. Springer, 2010, pp. 170–191.

[6] A. Z. Wyner and W. Peters, "On rule extraction from regulations." in *JURIX*, vol. 11, 2011, pp. 113–122.

[7] M. Grabmair, K. D. Ashley, R. Chen, P. Sureshkumar, C. Wang, E. Nyberg, and V. R. Walker, "Introducing LUIMA: An Experiment in Legal Conceptual Retrieval of Vaccine Injury Decisions Using a UIMA Type System and Tools," in *ICAIL '15: Proceedings of the 15th International Conference on Artificial Intelligence and Law*. New York, NY, USA: ACM, 2015, pp. 69–78.

[8] J. Šavelka, G. Trivedi, and K. D. Ashley, "Applying an Interactive Machine Learning Approach to Statutory Analysis," in *JURIX 2015*.

[9] L. Romano, M. Kouylekov, I. Szpektor, I. Dagan, and A. Lavelli, "Investigating a generic paraphrase-based approach for relation extraction," in *11th Conference of the European Chapter of the Association for Computational Linguistics*.

[10] V. Walker, J. Hae Han, X. Ni, and K. Yoseda, "Semantic Types for Computational Legal Reasoning: Propositional Connectives and Sentence Roles in the Veterans' Claims Dataset," in *ICAIL '17: Proceedings 2017*.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-21

Cloudy with a Chance of Concepts

Suzanne BARDELMEIJER[a], Alexander BOER[b], Radboud WINKELS[b]
[a] *Science Faculty, University of Amsterdam*
[b] *Leibniz Center for Law, University of Amsterdam*

Abstract. In this paper we present the results of a study to see whether automatically generated concept maps help users of legal information systems in understanding the topics of documents they retrieve. A small formative evaluation with novice users is presented. We did not find a significant difference between the ability to connect the correct visualisation to a document between a topic cloud and concept map approach. Topic clouds are probably a little easier to understand quickly in a superficial way.

Keywords: topic clouds, concept maps, LDA, legal recommender system

1. Introduction

Over the past few years, more and more legal documents have become publicly available online. In 2016 for instance, almost 34,000 new court decisions were published on the official Dutch judicial portal[1] [1]. These decisions are often complex due to their lengthy and complicated structure. A clear visualization of the different topics that a document deals with, might help in understanding a legal document quickly, and could help professionals and prevent novices from feeling overwhelmed by the length of the document. In this way it fits in our research on *legal recommender systems* in recent years (cf.[2]).

One way to visualize different topics is through topic clouds. A topic cloud is a visual representation of words (concepts) where the importance of a word in an underlying set of data (text) is expressed by its size [3].

Another way of visualizing knowledge in complex documents is via concept maps. A concept map is a graphical representation of knowledge in which the core concepts and the relations that connect these concepts are structured in a network diagram [4]. These concept maps, constructed by an expert in a certain field, help organize prior and newly acquired knowledge and therefore assist information gathering. In recent years, different algorithms have been created that quickly generate topic clouds automatically. However, the creation of a concept map is still a labour-intensive and time consuming process, mainly because substantial expert knowledge is typically needed.

With previous results and attempts in mind, this paper presents a new approach for the creation of concept maps from court decision documents. We aim to create small and comprehensive concept maps that capture the essence of document topics. We are interested in knowing to what extent it is possible to automatically create such a concept map from Dutch court decisions quickly. Furthermore, a comparison will be

[1] www.rechtspraak.nl

made between the generated concept maps and topic clouds to answer the question whether users have a preference for one or the other as a means of visualisation.

We will start with describing related work. Next, the concept map generation method is discussed, followed by the evaluation and interpretation of experimental results. To conclude this paper, recommendations for future research will be made.

2. Concept Maps & Topic Modelling

Fundamental research in concept mapping was first done by Novak and his researchers. Concept maps are graphical tools for representing knowledge through organizing concepts and their relations, to advance human learning and understanding [4]. The process of constructing a concept map normally requires a substantial amount of time and expert knowledge. A number of studies have focused on providing methods to help automate the process (e.g. [5]).

Most topic models are currently constructed using Latent Dirichlet Allocation, which was first presented by Blei e.a. [6]. The idea behind LDA is that documents are seen as a mixture of different topics, where a topic is formally defined as a distribution over a fixed vocabulary. LDA results in the creation of a topic model where each topic is associated with a document in different proportions and has probabilities of generating various words.

LDA operates under the bag-of-words assumption, meaning that word order is not taken into account. This assumption is plausible for the identification of a topic, but gives a disadvantage when interpreting them. Research showed that topic models based on LDA with a multi-word expression approach provide a better understanding for what a topic is about [7,8].

3. Research Method

We selected a sample of case law from the Dutch portal on immigration law for the period 2015-2017: 250 cases. Most Dutch court decisions have a common structure, starting off with a summary of the case, followed by the actual verdict which consists of the procedure, the considerations, and the decision. Since every section of the decision contains information about the case and could therefore be of interest in identifying underlying topics, all sections are considered relevant. All in all we processed 968 text files from the 250 cases. Every file contains a section of a court decision. A file is named after the case's ECLI number[2], which is a unique identification number, and a section number. To finalize the pre-processing step, we removed punctuation, and all capital letters were converted to lowercase. A list of stop words was created by computing the frequency of every word in the corpus. If a word occurred in more than 20% of all files, the word was added to the list of stop words.

To select promising n-grams, all text files were divided in bi-, tri- and tetra-grams. An n-gram is considered promising when none of its terms consist of one character or

[2] European Case Law Identifier (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2011:127:0001:0007:EN:PDF)

one digit. E.g. the bi-gram “is_a" is not considered promising. Also, none of the words should be a stop word. All promising n-grams are appended to the text file.

4. Building the Models

We used the open source implementation of LDA in MALLET[3] to build our topic models. We set the target number of topics to 50. Although the number of topics that LDA produces is arbitrary, it has a big influence on the informational value of the created topics. If the target number is reduced, separate topics have to merge; if the number increases, topics have to split. In order to ensure a topic is describing a theme that is well-represented in a set of documents, only documents that have more than 20% similarity with one of the topics are selected. To guarantee a topic is not too specific, the remaining topics are compared based on the number of documents they describe. For the 20 topics that describe most documents a concept map is created.

4.1. Creating Concept Maps

After finishing the steps above, 20 topics with their descriptive terms were extracted from the data set. These terms represent the concept nodes in a concept map of that topic. Since a concept map is created out of concepts and a limited number of most relevant relations between them, the next step is to assign weights to term pairs. If a topic is described by n terms, then there are $n(n\text{-}1)/2$ pairs, i.e. 171 pairs for 19 terms. The weight of a term pair is the summation, over the set of documents, of the products of the term frequencies of the term pair in the document. The weighted term pairs are stored in a list in descending order. The most informative pairs, with the highest weight, are selected as links and a concept map is constructed using CmapTools[4].
A comprehensible concept map should not contain concepts with more than three (incoming and outgoing) links. Therefore, once a concept already has three links with other concepts, new links to that concept are skipped. This process will continue until the concept map consists of 15 connected concepts. Figure 1 shows an example for topic 36 (in Dutch).

Although traditional concept maps exist of linking phrases containing a verb, the linking phrases in the created concept maps have a number that refers to a section of a court decision in which the link has the highest score. This approach ensures that the idea of the traditional concept map is maintained while avoiding natural language processing. In the evaluation this link information was not used. When integrated in an application, this information can provide users more insight in the origin of the link between two concepts.

4.2. Creating Topic Clouds

The output of LDA using MALLET is a table that expresses i.a. the importance of a term to a topic. This table gives valuable information; Some terms are more strongly connected to a topic than others.

[3] MAchine Learning for LanguagE Toolkit (http://mallet.cs.umass.edu)
[4] https://cmap.ihmc.us/

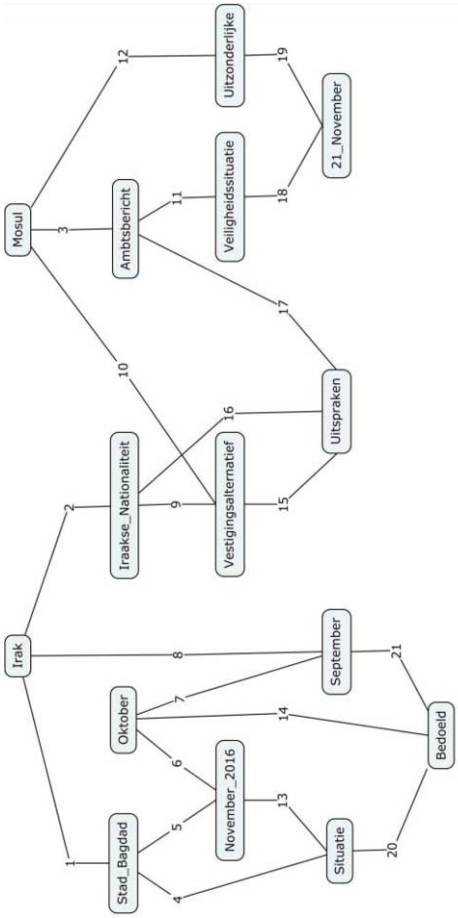


Figure 1: Example Concept Map for Topic 36

Figure 2: Example Topic Cloud for Topic 36

The strength of association of a term to a topic is expressed by the size of a term in a topic cloud. Figure 2 gives an example for topic 36 (in Dutch).

5. Formative Evaluation

We created concept maps and a topic clouds for each of the 20 topics that were extracted from our data. To evaluate these concept maps and topic clouds, six novices in the field of law participated in a study in which they had to complete sorting tasks. The study consisted of four tasks where in each task the participant was given a section of a case law and was asked to:

1. Rank in descending order four concept maps according to their similarity with the given document;
2. Rank in descending order four topic clouds according to their similarity with the given document.

Table 1 presents the documents, identified by their ECLI number, that were used for the study. These documents were selected since their content covers a wide variety of topics which makes them suitable for ranking. The topics beneath the ECLI number fit

the document in different degrees and this value decreases gradually. For instance, document ECLI:NL:RBDHA:2017:3176-2 is matched by topic 7 for 51%, topic 21 for 12%, etc.

The Spearman rank-order correlation, which is a non-parametric measure of the degree of association between two variables, was used to evaluate the survey. Table 2 presents the Spearman rank-order correlation coefficient and p-value for every task and in total. Although the number of participants is too small to draw any statistically sound conclusions, the correlation coefficients indicate a positive correlation for both concept maps and topic clouds. The values also suggest that perhaps participants were slightly more capable of adequate ranking using topic clouds than concept maps. Task 4 apparently either was more difficult than the other three tasks or the subjects were a bit tired and payed less attention.

Table 1: The four documents and their topics

ECLI:NL:RBDHA: 2017:3176-2		ECLI:NL:RBDHA: 2017:417-2	
Topic 7	0,513	Topic 48	0,295
Topic 21	0,118	Topic 33	0,113
Topic 33	0,056	Topic 30	0,067
Topic 16	0,014	Topic 12	0,015
ECLI:NL:RBDHA: 2017:2654-2		ECLI:NL:RBDHA: 2017:780-2	
Topic 36	0,489	Topic 10	0,328
Topic 29	0,131	Topic 45	0,106
Topic 3	0,021	Topic 30	0,030
Topic 9	0,009	Topic 12	0,015

Table 2: Spearman rank-order and p-values

	Concept Maps	p-value	Topic Clouds	p-value
1	1.00	0.0	1.00	0.0
2	0.77	1.00 E-05	0.97	0.0
3	0.93	0.0	1.00	0.0
4	0.40	0.053	0.39	0.056
Total	0.78	1.94 E-04	0.84	1.62 E-10

6. Conclusion

The method that is presented in this paper for the automatic creation of comprehensible concept maps from case law documents shows potential since novices in the field of law were able to rank the created concept maps adequately given a section of a case law. However, due to the marginal difference in performance of the two visualizations, we cannot conclude that there exists a distinct preference for either topic clouds or concept maps for the visualization of underlying topics in the case law documents.

While these results suggest that participants were able to make more correct rankings using topic clouds than concept maps, this does not necessary imply that topic clouds have more informational value. Participants could use two main methods to rank the concept maps and topic clouds. One method is aimed at identifying the internal structure of the given document, whereas the other is focused on words itself, as the LDA algorithm does. The first method requires substantial knowledge of the content of the document, while the second is based on superficial resemblance. Therefore, the results of the survey could be misleading if either the second method simply works better on LDA generated topic classifications, or if laymen use term frequency to determine similarity to both types of diagram. Further research could include experts as well to examine whether experts perform better or worse. In addition, obviously, more participants are needed in order to draw statistically sound conclusions.

Although a number of preliminary pre-processing steps were performed in order to establish a clear-cut topic model, this process can be improved. E.g. stemming was not performed since Dutch parsers do not achieve high results on Dutch case law documents due to their complicated structure and use of language. The development of parsers specially made for the analysis of legal documents could lead to better results in future research.

The last issue that needs to be addressed is the appearance of the concept maps. A traditional concept map is composed of a number of propositions in which two concepts are linked by a linking phrase. Although this linking phrase normally contains a verb phrase, the linking phrases in this case contain a whole section of a case law. Moreover, these sections corresponding to the linking phrases were not shown to participants during the evaluation. This deprived them of potentially valuable information about the origin of the link between two concepts and could therefore result in different ranks opposed to when this information was available to them. Further research should perhaps integrate the document sections corresponding to a linking phrase to provide this knowledge.

References

1. Rechtspraak, de (2017). *Jaarverslag 2016*. Retrieved from http://www.jaarverslagrechtspraak.nl ([Online; accessed 05-30-2017]). In Dutch.
2. Winkels, R., Boer, A., Vredebregt, B. and Someren, A. van. Towards a legal recommender system. In *Legal Knowledge and Information Systems: JURIX 2014*, the Twenty-seventh Annual Conference, pages 169-178. IOS Press, 2014.
3. Castano, S., Ferrara, A., & Montanelli, S. (2013). Mining topic clouds from social data. In *Proceedings of the fifth international conference on management of emergent digital ecosystems* (pp. 108{112). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2536146.2536171 doi: 10.1145/2536146.2536171
4. Novak, J., & Cañes, A. (2008). *The theory underlying concept maps and how to construct and use them*. Florida Institute for Human and Machine Cognition.
5. Villalon, J. & Calvo, R. (2009). Concept extraction from student essays, towards concept map mining. *IIEE International Conference on Advanced Learning Technologies*. DOI: 10.1109/ICALT.2009.215
6. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3* (Jan), 993{1022.
7. Blei, D. M., & Lafferty, J. D. (2009). Visualizing topics with multi-word expressions. arXiv preprint arXiv:0907.1013 .
8. Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. , 697-702. Retrieved from http://dx.doi.org/10.1109/ICDM.2007.86 doi: 10.1109/ICDM.2007.86.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-27

Dimensions and Values for Legal CBR

Trevor BENCH-CAPON, Katie ATKINSON
Department of Computer Science, The University of Liverpool, UK

Abstract. We build on two recent attempts to formalise reasoning with dimensions which effectively map dimensions into factors. These enable propositional reasoning, but sometimes a balance between dimensions needs to be struck, and to permit trade offs we need to keep the magnitudes and so reason more geometrically. We discuss dimensions and values, arguing that values can play several distinct roles, both explaining preferences between factors and indicating the purposes of the law.

Keywords. legal case based reasoning, dimensions, factors, values.

1. Introduction

Much work on reasoning with legal cases has been in terms of dimensions, introduced in HYPO [6], and factors, developed from dimensions in CATO [5]. Factors are stereotypical patterns of facts, either present or absent in a case, and, if present, favour either the plaintiff or the defendant. Dimensions, in contrast, are ranges of values (numeric or enumerated), running from an extreme pro-plaintiff point to an extreme pro-defendant point. The applicability of dimensions to a case, and the point at which the case lies, is determined by the case facts, and the dimension may favour either party. Most attention has been focussed on factors and factor based reasoning was formalised by Horty [16] and refined by Rigoni in [20]. A more detailed history of this line of development is given in [9]. More recently it has been argued that factors fail to capture some of the nuances present in legal Case Based Reasoning (CBR), and dimensions are needed to capture the *degree* to which a party is favoured [4] and to bridge from factors to the facts of a case (see[19] and [2]). This revival of interest has led to efforts by both Horty and Rigoni to extend their formalisations of factor based reasoning to dimensions in [15] and [21] respectively. Both Horty and Rigoni reduce dimensions to factors: in this paper will we retain magnitudes for some dimensions.

This paper is a shortened version of [7][1] and will focus on the main contributions of that paper. We represent domain knowledge as an Abstract Dialectical Framework (ADF) [13] as used in [3]. The key points are:

- Any legal CBR problem can be reduced to a series of steps involving at most two dimensions, so that higher dimensional spaces need not be considered;

[1]Available at http://intranet.csc.liv.ac.uk/research/techreports/tr2017/ulcs-17-004.pdf. For more context and detail see [7].

- The non-leaf nodes of the ADF can be seen as being one of five types, as determined by their children. For some nodes dimensions cannot be reduced to factors and need to retain their magnitude, to permit trade offs;
- Values are required to play several different roles, not just the expression of preferences.

After a summary of [15] and [21], we discuss each of these points in turn.

2. Formalising Factors and Dimensions

The formalisations of factor-based reasoning of both Horty and Rigoni are based on the method of expressing precedents as rules found in [18]. A case is considered to be a triple $\langle P, D, o\rangle$, where P is the set of all pro-plaintiff factors present in the case, D is the set of all pro-defendant factors present in the case and o is the outcome, either plaintiff (π) or defendant (δ). Now $P \rightarrow \pi$ will be the strongest reason to find for the plaintiff and $D \rightarrow \delta$ will be the strongest reason the find for the defendant. We can therefore deduce that either $P \succ D$ or $D \succ P$ depending on the value of o. A key insight of Horty is that $P \rightarrow \pi$ may be stronger than is required and some subset of P may be sufficient to defeat D. The use of P gives rise to what Horty terms the *rule* or *result* model and the subset the *reason* model.

In [15] Horty uses precedents to map a dimension into a factor. The point at which the factor becomes present depends on the facts of the case (result model) or the tests given in the opinion (reason model), as determined by the available precedents. However, as Horty shows, on this account the result and reason models collapse and the reason does not provide an effective constraint on subsequent decisions. Rigoni objects to these points and in [21] he avoids both of them by mapping a dimension into several factors similar to [19], with a point (the switching point (SP)) determined by the preferences at which the factors cease to favour one side and begin to favour the other. SP may lie on, or between, factors. Now reasons may be weaker than results in two ways: either they may contain fewer factors as in [15], or they may contain weaker factors from the same dimension. Rigoni also reconises that not all aspects of a case will contain magnitude and so cases are a four-tuple of *name*, *factors*, *dimensions* and *outcome*.

We regard Rigoni's account as improving on Horty's but claim that it cannot deal with questions of balance and trade off [17]. To handle this magnitudes need to be retained and the argumentation needs to become geometric as in [8] and [7]. With one dimension we can think in terms of *left* and *right* (or greater and less than), but with two dimensions we need to think in terms of *north-west* and *south-east* of the various points. The facts of the case and its result define an area where the decision must be followed, and the reason given offers a hypothetical set of facts which creates an area that presumptively favours the winning side. A new case may then fall into an area not yet covered by precedents and, depending on the outcome, will claim some of the space for the winning side. Figure 2 of [7] provides a relevant diagram.

3. How Many Dimensions Must we Consider?

In [8] the discussion was always in terms of two dimensions, but it was left open as to whether higher dimensional spaces might require consideration. In fact, just as any set of relations can be expressed in terms of binary relations and any k-SAT problem can be expressed as 3-SAT, it is possible to represent any domain so as to ensure that only two-dimensional spaces are needed. In [3] the ANGELIC methodology for representing domain knowledge as an Abstract Dialectical Framework (ADF) [13] was presented. Formally an ADF forms a three tuple: a set of nodes, a set of directed links joining pairs of nodes (a *parent* and its *children*), and a set of acceptance conditions to determine the status of the nodes. The nodes represent statements[2], which, in this context relate to issues, intermediate factors and base level factors, and acceptance conditions return a number between 0 and 1 representing the degree to which they favour the plaintiff. The links show which nodes are used to determine the acceptability of other nodes, so that the degree of a parent node is determined by its children. The acceptance condition for a node states how precisely its children relate to that node. In [1] it was shown that such an ADF could be rewritten as a *2-regular ADF*, in which every non-leaf node has at most two children. Since the degree to which a node favours the plaintiff depends only on its children this means that we need never consider more than two dimensions to resolve the acceptability of a node, and, since an ADF produced by the ANGELIC methodology forms a tree, the topmost node can be resolved without the need to consider more than two nodes at any given step.

4. Node Types

Like Rigoni, we recognise that not every aspect of a case requires representation of magnitude. In the original HYPO [6] there were thirteen dimensions. For two of these only one of the two extreme points was of interest; while for eight of them both end points were of interest, but not any intermediate points. One dimension was a set of enumerable points and the remaining two were continuous [9]. These four types represent a Horty style dimension, a pair of Rigoni-style factors, a Rigoni-style dimension and two irreducible dimensions requiring retention of magnitude, respectively. Interpreting these respectively as single factors, pairs of factors, sets of factors and dimensions, a given non-leaf node in our 2-regular ADF (leaf nodes are instantiated from the case facts) may have as children:

1. two factors;

[2]Contrary to the assertions of the reviewers of the original submission, these statements are not limited to two truth values. While originally in [13] they were presented as trivalent, later they were generalised in [14]: "In an ADF, an argument is either accepted (t), rejected (f), or undecided (u). We discuss how the ADF approach can be generalized to allow for more fine-grained distinctions. We consider acceptance degrees taken form an arbitrary domain of values possessing an adequate truth ordering and an information ordering. We show how to accommodate such values using an adequate characteristic operator. We illustrate the approach using degrees in the unit interval". Nor is there, *pace* the reviewers, any difficulty in connecting multi-valued statements with *AND* and *OR*. For example the techniques of fuzzy logic [22] could be used: this was the approach applied to ADFs in [10].

2. one dimension and one factor;
3. two dimensions;
4. one factor (the other child is a dummy node, for example, *true*);
5. one dimension (the other child may be a threshold, allowing the dimension to be coberted to a factor).

(1) is found in factor-based reasoning as formalised in [16] and [20]. In (2) the factor provides a *context* for the consideration of a dimension. In the fiscal domicile example discussed in [18], [15] and [21], citizenship may be a factor: if the person is a UK citizen a longer absence may be required before a change is made. Note that this aspect has no natural interpretation with magnitude: either one is a UK citizen or not. In (3) we have the kind of trade off mentioned above. The two dimensions describe points in a two dimensional space, and a line is drawn separating the area favouring one outcome from the area favouring the other outcome. Examples of (4) should be rare: the child can simply replace the parent. Finally in (5) we have a way of implementing thresholds. Thus the parent will be something like *sufficient absence*, and the purpose of the node is to provide a means of converting a dimension into a factor, much as envisaged by Horty in [16]. A set of such nodes, all with the actual point of the dimension as one child and a threshold as the other, would produce the set of factors envisaged in [21] and [19]. Thus only type (3) nodes will be resistant to the reduction to factors suggested by both [15] and [21], and require the style of reasoning of [8].

5. Relation with Values

Now we can reintroduce a relationship with purposes or values. The idea of values derives from [12] in which values were used to explain preferences between competing factors, and hence to resolve conflicts for which there was no precedent in terms of factors, as explained in detail in [11]. The existence of factors and dimensions in case law domains is justified by their role in enabling the consideration of the particular values the law is designed to promote. In [23] it was recognised that values might play two roles: justifying the presence of a rule, or justifying the inclusion of a particular antecedent in a rule.

In type (1) nodes, where the children are linked by AND, we ensure that both values are promoted, and where they are linked by OR we ensure that at least one of the values is promoted. Thus the role of nodes with two factors as children linked by AND or OR is to ensure that required values are given their due consideration. But there are also cases where the polarity of the two children are different: effectively the connective can be read UNLESS. This expresses a preference for the value associated with the exception. Note that only UNLESS requires a preference: AND or OR consider both values to be of importance.

The second kind of node is where we have a factor providing a context for a dimension. In the fiscal domicile example of [18] the length of absence might be considered differently for different types of citizen. UK citizens might require a longer absence than citizens of other countries who had been working here on a long term, but not permanent, posting. Thus we may envisage a parent *sufficient given citizenship*, with children *UK citizen* and *absence*. What we have here is

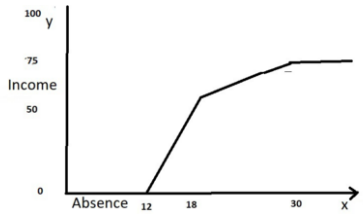


Figure 1. Possible trade off between absence and income percentage. The y-axis represents % income earned in UK, so that increasing values of y favour no change

effectively *two* distinct dimensions with different SPs. Which is used depends on whether the factor is present or not, and the applicable SPs will be specified in the acceptance conditions. The value served here is stability, but the context allows consideration of the value of mobility of labour, since we are allowing non-UK citizens an easier path to restoring their original fiscal domicile. Thus we are able to consider two values, or to consider what promotes a value in a particular context. Similarly nodes of type (5) allow consideration of what is sufficient to promote a value, but where the switching point at which the dimension becomes sufficient is the same for all cases. This permits a threshold for a factor to be determined by precedents, as envisaged in [15]: note that the different thresholds can be applied by using environment variables as antecedents in the acceptability conditions.

This leaves type (3), nodes with two genuine dimensions. Where they are linked by AND or OR, the role of values is the same as for two factors. For example, we can determine whether *both* sufficient absence (to promote stability) and a sufficient degree of engagement (shown by the percentage of foreign earnings, and promoting equity between countries) can be shown, so that the abstract factor *sufficient commitment* can be seen as present in the case. AND and OR, can be resolved using, for example, fuzzy logic style operators. Other type (3) nodes will be those where a balance needs to be struck (see [17]) and so there is a trade off between the dimensions. This is the situation considered in [8].

If we consider that the space can be divided by a single straight line we will have an equation of the form: $y = mx+c$ where m represents the slope of the line, and hence the degree of trade off. Very often, however, m will not be the same for all values of x: the amount of income required to trade off a year's absence, may change as absence increases. A fairly typical situation for absence and income percentage in the fiscal domicile example is shown in Figure 1.

In Figure 1 we have a minimum absence and a minimum percentage of income, with two different rates of trade off in between. To describe this we need a set of four equations covering the various ranges: $0 \leq x \leq 12 : y = 0$, $12 \leq x \leq 18 : y = mx$, $18 \leq x \leq 30 : y = nx$ (where $n < m$) and $x \geq 30 : y = 75$.

The coefficient of x is important because it represents the degree of trade off, the relative weight to be given to the different values at different points. In Figure 1 we have sharp changes of slope, represented by a set of line segments, but often there is a gradual and regular change, better represented by a curve rather than a set of straight line segments, with the gradient varying as a function of x. This function will determine whether the curve becomes increasingly steep or

increasingly shallow. In some cases we can imagine the curve changing direction entirely: for a discussion of this point see [7].

In conclusion, we have discussed how we can extend [15] and [21] with dimensions which cannot be reduced to sets of factors.

References

[1] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. Factors, issues and values: Revisiting reasoning with cases. In *Proceedings of the 15th International Conference on AI and Law*, pages 3–12. ACM, 2015.

[2] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. Angelic secrets: bridging from factors to facts in US Trade Secrets. In *JURIX 2016*, pages 113–118. IOS Press, 2016.

[3] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. A methodology for designing systems to reason with legal cases using abstract dialectical frameworks. *AI and Law*, 24(1):1–49, 2016.

[4] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. Statement types in legal argument. In *Proceedings of JURIX 2016*, pages 3–12. IOS Press, 2016.

[5] V Aleven. Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment. *Art. Int.*, 150(1-2):183–237, 2003.

[6] K Ashley. *Modeling legal arguments: Reasoning with cases and hypotheticals*. MIT press, Cambridge, Mass., 1990.

[7] K Atkinson and T Bench-Capon. *Dimensions and Values for Reasoning with Legal Cases*. Technical Report ULCS 17-004, Dept of Computer Science, Univ of Liverpool, 2017.

[8] T Bench-Capon. Arguing with dimensions in legal cases. In *Proceedings of CMNA 2017*, pages 1–5, 2017.

[9] T Bench-Capon. Hypo's legacy: introduction to the virtual special issue. *AI and Law*, 25(2):205–250, 2017.

[10] T Bench-Capon and T Gordon. *Tools for Rapid Prototyping of Legal Case Based Reasoning*. Tech Report ULCS 15-005, Dept of Computer Science, Univ of Liverpool, 2017.

[11] T Bench-Capon and G Sartor. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150(1-2):97–143, 2003.

[12] D Berman and C Hafner. Representing teleological structure in case-based legal reasoning: The missing link. In *Proceedings of the 4th ICAIL*, pages 50–59, 1993.

[13] G Brewka, S Ellmauthaler, H Strass, J Wallner, and P Woltran. Abstract dialectical frameworks revisited. In *Proceedings of the Twenty-Third IJCAI*, pages 803–809, 2013.

[14] Gerhard Brewka. Weighted abstract dialectical frameworks. In *Workshop on Argument Strength*, page 9, 2016.

[15] J Horty. Reasoning with dimensions and magnitudes. In *Proceedings of the 16th ICAIL*, pages 109–118. ACM, 2017.

[16] J Horty and T Bench-Capon. A factor-based definition of precedential constraint. *AI and Law*, 20(2):181–214, 2012.

[17] M Lauritsen. On balance. *AI and Law*, 23(1):23–42, 2015.

[18] H Prakken and G Sartor. Modelling reasoning with precedents in a formal dialogue game. *AI and Law*, 6(3-4):231–87, 1998.

[19] H Prakken, A Wyner, T Bench-Capon, and K Atkinson. A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. *Journal of Logic and Computation*, 25(5):1141–1166, 2015.

[20] A Rigoni. An improved factor based approach to precedential constraint. *AI and Law*, 23(2):133–160, 2015.

[21] A Rigoni. Representing dimensions within the reason model of precedent. *AI and Law*, page In Press: Available online October 2017, 2018.

[22] L. A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[23] T Zurek and M Araszkiewicz. Modeling teleological interpretation. In *Proceedings of the 14th ICAIL*, pages 160–168. ACM, 2013.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-33

Timed Contract Compliance Under Event Timing Uncertainty

María-Emilia CAMBRONERO [a] and Luis LLANA [b] and Gordon J. PACE [c,1]
[a] *Department of Computer Science, University of Castilla-La Mancha, Albacete, Spain*
[b] *Department of Computer Science and Computation, Complutense | University of Madrid, Spain*
[c] *Department of Computer Science, University of Malta*

Abstract. Despite that many real-life contracts include time constraints, for instance explicitly specifying deadlines by when to perform actions, or for how long certain behaviour is prohibited, the literature formalising such notions is surprisingly sparse. Furthermore, one of the major challenges is that compliance is typically computed with respect to timed event traces with event timestamps assumed to be perfect. In this paper we present an approach for evaluating compliance under the effect of imperfect timing information, giving a semantics to analyse contract violation likelihood.

Keywords. Contract compliance, Formal semantics, Real-time contracts, Fuzzy time

1. Introduction

Many real-life contracts include concrete time constraints, whether placing limits by when obligations have to be discharged e.g. *"Money is to be made available to the client with 48 hours of a request for redemption"*, or whether it identifies a time window during which an event is prohibited e.g. *"Once disabled, a user may not log in for 1 hour"* or even through temporal delays e.g. *"After accessing the service for 30 minutes, the user is obliged to pay within 5 minutes or lose the right to use the service further"*. A number of contract languages which allow for the description of such real-time matters have been proposed in the literature, including ones based on deontic logic e.g. [1,2] and automata e.g. [3]. However, one common feature of these formal approaches is that they handle compliance analysis in a *crisp* manner — for a given contract and a sequence of timed events (each carrying a timestamp), they enable the identification of whether or not that trace violates the contract, giving a *yes* or *no* answer.

Consider the contract clause which states that $C \stackrel{df}{=}$ *"Once disabled, a user may not log in for 1 hour"*, and the following event trace:

$$tr \stackrel{df}{=} \langle (login, 02h24m58s), (disable, 02h25m02s), (login, 03h25m01s) \rangle$$

Typically, the analysis would deduce that contract C has been violated by trace tr due to the second *login* event happening within less than an hour of the *disable* event. The major

[1]Corresponding Author: Gordon J. Pace; E-mail: gordon.pace@um.edu.mt.

issue with such an analysis is the difficulty of obtaining perfect timestamps, particularly if the sources of events may occur in different locations.

In this paper we outline our initial attempts at adapting techniques originally developed for real-time logics and calculi [13] to enable compliance analysis starting from traces with fuzzy timed events i.e. the timestamp for each event is *not* a single point in time, but a function over time which indicates the likelihood of the event having happened at that point in time.

2. Fuzzy Timed Event Traces

Trace-based semantics of contracts define whether for a particular contract a given trace of events violates that contract or not. Given an alphabet of event names which can be observed EVENT, such semantics typically take a trace ranging over EVENT^*. When contracts refer to real-time, the traces have to be augmented to have every event tagged by a timestamp indicating when it occurred. Taking time to range over the non-negative real numbers: $\text{TIME} \stackrel{df}{=} \mathbb{R}_0^+$, a real-time trace tr ranges over sequences of timed events $tr \in (\text{EVENT} \times \text{TIME})^*$ (with the assumption that time is monotonically increasing along the trace), or just using a set of timed events $tr \in 2^{\text{EVENT} \times \text{TIME}}$ (since the timestamps implicitly indicate the ordering).

In our case, the time stamps are no longer exact point values, and we assume that we can only give a likelihood of an event having happened at a particular time. Thus, rather that associating every observed event with a single value over TIME, we will use an approach from fuzzy logic, giving a time distribution, associating every time value with the likelihood of the event having happened at that point in time i.e. $\text{TIMEDISTRIBUTION} \stackrel{df}{=} \text{TIME} \to [0,1]$.

Definition 1 *A* fuzzy-timed observation *is a pair* $\langle a, T \rangle \in \text{FUZZYOBSERVATION}$*, where* $a \in \text{EVENT}$ *is the event name, and* $T \in \text{TIMEDISTRIBUTION}$ *is the timing distribution of that event. A* fuzzy-timed trace $tr \in \text{FUZZYTRACE}$ *is a pair* $\langle es, \Delta \rangle$*, consisting of (i)* $es \in \mathscr{M}(\text{FUZZYOBSERVATION})$*, a finite multiset of fuzzy-timed observations*[2]*; and (ii)* $\Delta \in \text{TIME}$*, the event horizon indicating that the events recorded are from the initial time window from time 0 and* Δ *(i.e. events happening beyond this time window are not recorded).*

It is worth noting that the imprecision inherent in the traces is limited to the timing of the events. We assume that the multiset of event names recorded is faithful with respect to what really happened i.e. (i) all events are recorded (no events are lost); (ii) events are not wrongly observed (event integrity); and (iii) no extraneous events are inserted (no spontaneous generation of event).

Also note, that if the fuzzy observation distributions are probabilistic ones, and independent of each other, then we can use a probabilistic approach (e.g. if we are given two fuzzy-timed observations $\langle a_1, T_1 \rangle$ and $\langle a_2, T_2 \rangle$, then the probability of both events hap-

[2]We use the notation $\mathscr{M}(X)$ to denote multisets with elements from X. Note that a sequence of fuzzy-timed observations cannot be used, since there is now no canonical ordering of events, and neither is a set of observations sufficient, since we may observe two events with the same name and with the same distribution function.

pening at time t would be $T_1(t) \times T_2(t)$). However, since this independence is not always easy to guarantee, we adopt a fuzzy logic approach and will combine likelihood values using generic binary operators $\otimes$ (the likelihood of the two observations to happen, a *triangular norm* [11]) and $\oplus$ (the likelihood of either of the two observations to happen, a *triangular conorm*). We will write $\prod$ and $\sum$ for the generalised versions of $\otimes$ and $\oplus$.

We will define a number of operations on fuzzy observations and traces to be used in the rest of the paper.

Definition 2 *We will write eventHorizon(tr) to refer to the event horizon of trace tr i.e. eventHorizon$(\langle es, \Delta \rangle) \stackrel{df}{=} \Delta$. Fuzzy-timed observations and fuzzy-timed traces can be shifted earlier in time using the* time shift *operator $\ll$:*

$$\langle a, T \rangle \ll \delta \stackrel{df}{=} \langle a, \lambda t.T(t \ominus \delta) \rangle$$
$$\langle es, \Delta \rangle \ll \delta \stackrel{df}{=} \langle \{e \ll \delta \mid e \in es\},\ \Delta \ominus \delta \rangle$$

where $x \ominus y \stackrel{df}{=} \max(x - y, 0)$.

We define the function occurances, which given an event and a fuzzy-timed trace, returns a multiset of all time distributions which may occur according to the given trace:

$$occurances_a(\langle es, \Delta \rangle) \stackrel{df}{=} \{T \in \text{TIMEDISTRIBUTION} \mid \langle a, T \rangle \in es\}$$

Finally, we define the likelihood that for a given fuzzy-timed trace es, event a has not happened in the initial time window $[0, \delta]$, written $absence_{es}(a, \delta)$ as follows:

$$absence_{tr}(a, \delta) \stackrel{df}{=} \prod_{T \in occurances_a(tr)} 1 - \int_0^{\delta} T(t)\, dt$$

3. A Timed-Contract Language

In order to define compliance and violation of fuzzy-timed traces, we will take a real-time deontic logic covering obligations, prohibitions, recursion and reparations to show how typical deontic operators can be given a trace semantics under imprecisely timed observations. The syntax of the real-time deontic logic is the following:

$$\mathscr{C} ::= \top \mid \bot \mid wait_\delta(\mathscr{C}) \mid \mathscr{O}_{\leq \text{TIME}}(\text{EVENT})(\mathscr{C}, \mathscr{C}) \mid \mathscr{F}_{\leq \text{TIME}}(\text{EVENT})(\mathscr{C}, \mathscr{C}) \mid \mu X.\mathscr{C} \mid X$$

The core of the calculus are obligations and prohibitions, written as $\mathscr{O}_{\leq \delta}(a)(C_1, C_2)$ and $\mathscr{F}_{\leq \delta}(a)(C_1, C_2)$ respectively. Obligation $\mathscr{O}_{\leq \delta}(a)(C_1, C_2)$ indicates that event a is to be performed before δ time units pass. If a is performed before the deadline, the continuation contract C_1 starts being enforced, but if a is not performed within δ time units, the reparation contract C_2 instead starts being enforced. Dually, prohibition $\mathscr{F}_{\leq \delta}(a)(C_1, C_2)$ indicates that event a is prohibited for the upcoming δ time units. If a occurs in this time frame, the reparation contract C_2 is triggered, but if it does not, then the continuation contract C_1 starts being enforced instead. The fact that we give both obligation and prohibition modalities a continuation and reparation, the two modalities are direct duals of each other: $\mathscr{F}_{\leq \delta}(a)(C_1, C_2)$ yields the same top-level violations (i.e. ignoring violations for which a reparation is defined) as $\mathscr{O}_{\leq \delta}(a)(C_2, C_1)$.

The contract $wait_\delta(C)$ acts like contract C, but starting after δ time units have elapsed. The base contracts $\top$ and $\bot$ are used to denote the contracts which are, respectively, immediately satisfied and violated. Finally, the μ operator is used to denote re-

cursion — such that the contract $\mu X.C$ will act like contract C except that every free instance of X in C will act like $\mu X.C$ itself.

Note that, for simplicity, all obligations and prohibitions have continuations and reparations, but if these are not desired, one can use the base contracts $\top$ and $\bot$. For example, to state that one is obliged to logout in 30 minutes, with no reparation or continuation, one would write: $\mathscr{O}_{\leq 30}(logout)(\top,\bot)$. In the rest of the paper, to avoid syntactic overload we will leave out the $\top$ and $\bot$ continuation and reparation e.g. simply writing $\mathscr{O}_{\leq 30}(logout)$.

We will now give a fuzzy-timed trace semantics to the timed contract logic. It is worth observing that when giving a trace semantics for crisp observations, one would typically define a (crisp) relation between traces and contracts such that a trace and contract are related if and only if the trace led to a violation of the contract. In contrast, in the case of fuzzy-timed traces, such a relation can only provide fuzzy information — i.e. we will have a functions $[\![C]\!]^{tr}_{vio}$ indicating the likelihood of fuzzy-timed trace *tr* violating contract C.

We can define the semantics of the timed contract logic with respect to a fuzzy-timed trace in this manner. As the base case for the semantics, we can assert that a trace with an event horizon of 0 cannot result in a violation. Trivially violated and satisfied contracts similarly given certain results (1 and 0 respectively), while a contract starting with a *wait* clause simply shifts the timestamps of the trace and analyses the resulting trace with the continuation of the contract. Obligation is the most complex operator, for which we have to separately compute whether the obliged action is performed or not, and combine with the continuation and reparation of the obligation. Prohibition is given a semantics in terms of obligation, while the semantics of recursion uses unrolling of the definition.

Definition 3 *The trace semantics of violation are defined in terms of the violation function* $[\![-]\!]^{-}_{vio} \in \mathscr{C} \times \textsc{FuzzyTrace} \to [0,1]$*, such that for a given contract C and fuzzy trace tr,* $[\![C]\!]^{tr}_{vio}$ *gives the likelihood of the the observations in trace tr violating contract C, and is defined as follows:*

If $eventHorizon(tr) = 0$*:*

$$[\![C]\!]^{tr}_{vio} \stackrel{df}{=} 0$$

Otherwise:

$$[\![\top]\!]^{tr}_{vio} \stackrel{df}{=} 0$$

$$[\![\bot]\!]^{tr}_{vio} \stackrel{df}{=} 1$$

$$[\![wait_{\delta}(C)]\!]^{tr}_{vio} \stackrel{df}{=} [\![C]\!]^{tr \ll \delta}_{vio}$$

$$[\![\mathscr{O}_{\leq\delta}(a)(C_1,C_2)]\!]^{tr}_{vio} \stackrel{df}{=} \sum_{T \in occurances_a(tr)} \int_0^{\delta} absence_{tr}(a,t) \otimes T(t) \otimes [\![C_1]\!]^{tr \setminus \{(a,T)\} \ll t}_{vio} \, dt \oplus absence_{tr}(a,\delta) \otimes [\![C_2]\!]^{tr \ll \delta}_{vio}$$

$$[\![\mathscr{F}_{\leq\delta}(a)(C_1,C_2)]\!]^{tr}_{vio} \stackrel{df}{=} [\![\mathscr{O}_{\leq\delta}(a)(C_1,C_2)]\!]^{tr}_{vio}$$

$$[\![\mu X.C]\!]^{tr}_{vio} \stackrel{df}{=} [\![C[X \setminus \mu X.C]]\!]^{tr}_{vio}$$

Provided that all uses or recursion are guarded (i.e. the recursion variable occurs after an obligation, prohibition or wait), the finite event horizon, and the finite size of

the timed observations recorded in the trace guarantee termination of recursion, thus ensuring that the semantics are well-defined.

4. Related Work

In the literature the use of fuzzy logic approaches for contracts are typically limited to analysing possible observational continuations, e.g. computing the likelihood of future failure given what has already been observed e.g. [10]. Even when encoded within the logic, most work deals with a discrete time model. For instance, Figeri *et al.* [4] present a temporal logic Fuzzy-time Temporal Logic (FTL), to express the temporal imprecision allowing to express vague temporal notions such as *soon*. Crespo *et al.* [13] present another work considering the extension of time constraints with fuzzy methods for timed automata, while Alur *et al.* [14], extend timed automata with perturbed clocks.

In practically all these works discussed, it is worth noting that the fuzziness is typically dealt with at the logic level — the specification language or logic allows for the expression of fuzzy notions of time. Our approach takes the dual view, and assigns fuzziness to the timing of the observations. In terms of expressivity, the two approach seem to be equally expressive. However, we believe that our approach is more appropriate in a deontic setting. For instance, consider trying to regulate a speed limit of 30km/h in a particular area. In order to enforce such a regulation, cameras are used, with imprecise timers. Because of this imprecision, the police may decide to prosecute only those who were observed driving at 40km/h or faster, since even taking into account the timing imprecision, it can be ascertained that the speed limit was exceeded. If, however, the cameras are replaced with more accurate ones, it may become viable to (fairly) prosecute those exceeding just 35km/h. If we were to take the approach that fuzzy timing should appear in the regulation itself, one would have to update the regulations whenever a camera is changed to a more (or less) accurate one. In contrast, with our approach, the regulation remains unchanged, *"You may not exceed 30km/h"*, but the uncertainty distributions in the observations allow the calculation of the probability or likelihood of an observed car to have actually been overspeeding.

5. Conclusions

We have proposed a fuzzy time trace semantics for violations of timed contracts. The approach allows the factoring in of imprecise measurements when recording timestamps of events (e.g. due to communication lag or due to unsynchronised clocks) while still allowing the calculation of the likelihood of a violation of the contract actually having taken place. In contrast to specification languages, where observational error is typically encoded with the property or specification, in a deontic setting, we would like to keep a canonical form of the regulating text, and factor in the error in the input trace. In practice, such semantics can be used, for instance, to regulate financial transactions, where the distributed nature of the interacting subsystems (account holder, receiver of funds, node logging events, etc.) means that precise timing of events is virtually impossible.

As it stands, the work has a number of limitations which we are currently addressing. On one hand, we would like to extend the timed deontic logic to include conjunction and

choice over contracts, thus widening its expressivity. Furthermore, the approach we have presented in this paper places no constraint on the form of the time-stamp distribution functions, resulting in the semantics being of limited practical use due to difficulty in computing them. We are, however, exploring limiting these functions (e.g. limiting time-stamp distribution functions to trapezoidal shaped ones), in order to be able to compute the results of the semantics automatically.

References

[1] Jan M. Broersen, Frank Dignum, Virginia Dignum, and John-Jules Ch. Meyer. Designing a deontic logic of deadlines. In *7th International Workshop on Deontic Logic in Computer Science, DEON 2004, Madeira, Portugal, May 26-28, 2004. Proceedings*, pages 43–56, 2004.

[2] María Emilia Cambronero, Luis Llana, and Gordon J. Pace. A timed contract-calculus. Technical Report CS-2017-02, Department of Computer Science, University of Malta, 2017.

[3] Enrique Martínez, María-Emilia Cambronero, Gregorio Díaz, and Gerardo Schneider. Timed automata semantics for visual e-contracts. In *Proceedings Fifth Workshop on Formal Languages and Analysis of Contract-Oriented Software, FLACOS 2011, Málaga, Spain, 22nd and 23rd September 2011*, 2011.

[4] Achille Frigeri, Liliana Pasquale, and Paola Spoletini. Fuzzy time in LTL. *CoRR*, abs/1203.6278, 2012.

[5] Martin Leucker and César Sánchez. *Regular Linear Temporal Logic*, pages 291–305. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[6] Joaquin Perez, Jaime Jimenez, Asier Rabanal, Armando Astarloa, and Jesús Lázaro. Ftl-cfree: A fuzzy real-time language for runtime verification. *IEEE Trans. Industrial Informatics*, 10(3):1670–1683, 2014.

[7] Barbara Pernici and Seyed Hossein Siadat. Selection of service adaptation strategies based on fuzzy logic. In *SERVICES*, pages 99–106, 2011.

[8] Anderson Francisco Talon and Edmundo Roberto Mauro Madeira. Comparison between light-weight and heavy-weight monitoring in a web services fuzzy architecture. *Procedia Computer Science*, 64 (Complete):862–869, 2015.

[9] Anderson Francisco Talon and Edmundo Roberto Mauro Madeira. Improvement of e-contracts accomplishments by self-adaptive fuzzy architecture. In *2015 IEEE International Conference on Services Computing, SCC 2015, New York City, USA, 2015*, pages 507–514, 2015.

[10] Anderson Francisco Talon and Edmundo Roberto Mauro Madeira. A fuzzy scheduling mechanism for a self-adaptive web services architecture. In *ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems, Volume 1, Porto, Portugal, April 26-29, 2017*, pages 529–536, 2017.

[11] E.P. Klement, Radko Mesiar and Endre Pap. Triangular Norms. Springer Netherlands, 2000.

[12] Anderson Francisco Talon, Edmundo Roberto Mauro Madeira, and Maria Beatriz Felgar de Toledo. Self-adaptive fuzzy architecture to predict and decrease e-contract violations. In *2014 Brazilian Conference on Intelligent Systems, BRACIS 2014, Sao Paulo, Brazil, October 18-22, 2014*, pages 294–299, 2014.

[13] F. Javier Crespo, Alberto de la Encina and Luis Llana, Fuzzy-Timed Automata, in the *Proceedings of Formal Techniques for Distributed Systems 2010,* Springer Berlin Heidelberg 2010.

[14] Rajeev Alur, Salvatore La Torre, and P. Madhusudan, Perturbed Timed Automata, in *Lecture Notes in Computer Science: Hybrid Systems: Computation and Control 3414,* 2005.

[15] Rajeev Alur, Salvatore La Torre, and P. Madhusudan. Perturbed timed automata. In *Hybrid Systems: Computation and Control, 8th International Workshop, HSCC 2005, Zurich, Switzerland, March 9-11, 2005, Proceedings*, pages 70–85, 2005.

[16] Alexandre Donzé and Oded Maler. Robust satisfaction of temporal logic over real-valued signals. In *Formal Modeling and Analysis of Timed Systems - 8th International Conference, FORMATS 2010, Klosterneuburg, Austria, September 8-10, 2010. Proceedings*, pages 92–106, 2010.

[17] Georgios E. Fainekos and George J. Pappas. Robustness of temporal logic specifications for continuous-time signals. *Theoretical Computer Science*, 410(42):4262 – 4291, 2009.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-39

Detecting Agent Mentions in U.S. Court Decisions

Jaromír ŠAVELKA [a,b,1], and Kevin D. ASHLEY [a,b,c]
[a] *Intelligent Systems Program, University of Pittsburgh*
[b] *Learning Research and Development Center, University of Pittsburgh*
[c] *School of Law, University of Pittsburgh*

Abstract. Case law analysis is a significant component of research on almost any legal issue and understanding which agents are involved and mentioned in a decision is integral part of the analysis. In this paper we present a first experiment in detecting mentions of different agents in court decisions automatically. We defined a light-weight and easily extensible hierarchy of agents that play important roles in the decisions. We used the types from the hierarchy to annotate a corpus of US court decisions. The resulting data set enabled us to test the hypothesis that the mentions of agents in the decisions could be detected automatically. Conditional random fields models trained on the data set were shown to be very promising in this respect. To support research in automatic case-law analysis we release the agent mentions data set with this paper.

Keywords. case law, legal analysis, agent mentions, named entity recognition, conditional random fields

Introduction

We examine the possibility of automatic detection of agent mentions in case law analysis. This would be an important prerequisite for many applications, such as attribution resolution. It may also become an important component of other applications such as information retrieval or summarization. We assess the hypothesis that a simple sequential model that uses low-level textual features could learn to detect agent mentions automatically (hypothesis 1). Obtaining data for a statistical learning model is expensive. Therefore we explore the relatedness of the task when performed on different areas of law (cyber crime and intellectual property). We first confirm that when a model is trained on decisions from one area and applied to texts from the other domain the performance is lower (hypothesis 2). But we also show that using texts from multiple domains may lead to higher quality predictive models (hypothesis 3).

1. Background and Motivation

Case law analysis is the process of determining which prior court decisions apply to a case, how they apply, and the effect of this application. In the context of judicial decision-

[1]Corresponding Author: Jaromír Šavelka, Learning Research and Development Center, 3939 O'Hara St, Pittsburgh, PA 15260, USA; E-mail: jas438@pitt.edu.

making the objective of the analysis could be to generate persuasive case-based arguments. These arguments could play a pivotal role in how a court decides a case. In the American legal system under the common law doctrine of *stare decisis*, like cases are decided alike. [3, p. 9]

Case law analysis encompasses two different, yet closely related, activities. First, a lawyer needs to *identify a set of decisions* that are relevant to argumentation in the given case. Then, from the texts of the decisions one *extracts valuable information* such as: authoritative applications of the rule conditions and concepts to identified situations, a ground truth for testing predictions about outcomes in new cases with new evidence, patterns for successful and unsuccessful argumentation, and guidance in retrieving, extracting, and organizing evidence for new arguments and new situations. [5, p. 176] This is an iterative process where the newly found pieces of information inform search for additional relevant decisions.

Existing legal information retrieval (IR) systems are relatively well suited to support the task of identifying relevant decisions. By means of a search query a lawyer specifies a hypothesis about what words and phrases are likely to occur in relevant decisions. The IR systems are much less equipped to help with the extraction of valuable pieces of information from the texts. Most of the times this needs to be done manually.

It has been extensively argued and shown that computational support for directly retrieving arguments and argument-related information (AR) would be extremely valuable. [4,9] Despite the great promise there is still a considerable gap between the demonstrated automatic analysis capabilities and a full-blown AR system. [2] Due to peculiarities of legal texts even the most foundational natural language processing (NLP) techniques are often performed poorly. One such technique is the detection of agent mentions. Being able to recognize when an agent is mentioned is vital, among many other applications, for attribution resolution. [16] This is why we focus on the capability to detect agent mentions automatically.

2. Task Definition, Proposed Solution, and Working Hypothesis

Detecting agent mentions amounts to recognizing when a word or a phrase denotes an agent. An agent could be any person or organization from informal groups to business companies and governmental entities. As it turns out a typical court decision contains many mentions of agents as shown in the following example:

> The magistrate judge denied the second motion to compel because Mavrix failed to notify the anonymous parties of the pending motion. Mavrix moved the district court for review of the magistrate judge's order, which the district court denied on the basis of the moderators' First Amendment right to anonymous internet speech.

In the short excerpt above there are multiple mentions of a judge, Mavrix (a party), anonymous parties, a court, and moderators—all of these are agents. Since we aim for the maximum possible coverage even words such as possessive adjectives (e.g., his, their) are considered agent mentions.

Apart from recognizing that an agent is being mentioned it would be very useful to understand what kind of agent it is. This is especially true for agents that play specific roles in a case (e.g., a court, a party, or a witness). For this reason we defined a

Agent		
Person		Organization
Attorney	Party	Jury
Judge	Amicus Curiae	Legislator
Expert		Court
Witness		

Table 1. The light-weight 3-level agent types hierarchy. The top-level type Agent is differentiated into the Person and the Organization types. These are further distinguished in the bottom level.

light-weight and easily extensible hierarchy of agents. The hierarchy is schematically depicted in Table 1. Different types of agents are organized into three layers. On the top level there is the Agent type that divides into the Person and Organization types (middle level). These two types are further differentiated into the Judge, Party, Attorney, Witness, Expert, Court, Jury, Amicus Curiae, and Legislator types (bottom level).

The task of detecting agent mentions in texts of the court decisions can be understood in the following way: 1. find all the text spans denoting agents; 2. Classify each such text span with the most appropriate agent type from the hierarchy. We hypothesize that both steps of this task could be performed automatically using a sufficiently well trained sequence labeling model such as conditional random fields (CRF).

We expect that the task depends on the domain, that is the area of legal regulation such as cybercrime or copyright. Intuitively, agent mentions such as "a victim" or "an investigator" are more likely to appear in a cybercrime decision whereas "a copyright holder" would more often appear in a copyright case. We also expect the domains to be related in a sense that some knowledge about detecting the mentions in one domain would be useful in a different one.

3. Related Work

Peters and Wyner [13] underscore the importance of identifying agents in legal documents: "At a more fine-grained level, it is important to access who bears what role with respect to the norm, that is, who is the responsible agent or the receiving party within the action." They employed a combination of pattern-identifying rules, parsing and semantic information about verbs and their arguments as heuristics to identify role bearing agents in European Directives. Similarly, the xmLegesExtractor tool used knowledge-engineered text classification rules and natural language parsing to extract role-playing agents regarding a statutory duty such as addressee, action, and counter-party. [8]

Researchers have also applied supervised machine learning to extract relevant functional elements from multiple states' statutes dealing with public health emergencies including the types of public health system participants who are the acting and receiving agents of regulatory directives. [15] The information is used among other things to construct statutory network diagrams with which to compare the states' regulatory schemes.

Quaresma and Goncalves [14] used parsing for named entity recognition of organizations in a corpus of international agreements from the Euro-Lex site and machine learning to identify types of agreement. The intention was to enrich an ontological index for improving information retrieval.

According to Faiz and Mercer [7] "extraction of many higher order relations is dependent on coreference resolution. ... [A]ugmenting a coreference resolution module in [a] pipeline would be an immediate improvement." For instance, a robust ability to iden-

	# of docs	# of chars	# of tokens	# of sentences	longest	average	shortest
cyber-crime	5	199980	71100	1772	61703 (c)	39996.0 (c)	28306 (c)
					20881 (t)	14220.0 (t)	10414 (t)
					513 (s)	354.4 (s)	250 (s)
intellectual-property	5	247042	90286	2084	75625 (c)	49408.4 (c)	36823 (c)
					27915 (t)	18057.2 (t)	13144 (t)
					729 (s)	416.8 (s)	291 (s)

Table 2. The data set summary statistics. In the last three columns the length is reported in characters (c), tokens (t), and sentences (s).

tify agents referred to in legal decisions is necessary to deal with the problem of attribution, "determining who believes a stated proposition to be true." [17] As Walker argues, "accurate attribution can be a critical task for argumentation mining." For example, it can help to assign legal sentence role types in an annotation pipeline "by distinguishing among ... the testimony of an expert witness, ... or a conclusion or finding of fact by the judge." Automatically detecting distinctions such as between evidence statements and a court's findings of fact could help transform legal IR into argument retrieval. [4]

Some research has focused on identifying agent references in legal decisions. Dozier et al. [6] applied a combination of table lookup, contextual rules, and a statistical model (CRF) to recognize types of entities in captioned legal decisions including jurisdiction, court, and judge. In order to resolve the entities of various types, a SVM model learned to match the extracted entity types and information against authoritative files of actual jurisdictions, courts, and judges.

Al-Kofahi, et al. [1] presented an algorithmic technique that combined parsing, domain knowledge about court hierarchies, and discourse analysis to identify treatment history language in court opinions. Such language includes references to courts as agents as in, "The court in Jones held that ... On the other hand, the district court of Oklahoma, held that ..."

4. Agent Mentions Data Set

We downloaded ten court decisions from the online Court Listener[2] and Google Scholar services.[3] Five of these decisions are from the area of cyber crime (cyber bullying, credit card frauds, possession of electronic child pornography), and five cases involve intellectual property (copyright, trade marks, patents). Detailed information about the texts is provided in Table 2. We use cases from the two different areas of law to measure how well the trained models generalize. We also explore if a model trained for one area of law could improve the performance of a model trained for a different domain.

We created guidelines for manual annotation[4] of the decisions with the types from the hierarchy introduced in Section 2. The two human annotators (the authors) were instructed to aim for the:

1. *Full coverage* – every single word or a phrase that denotes an agent should be annotated with one of the available types.

[2] `www.courtlistener.com`

[3] `scholar.google.com`

[4] Accessible at `luima.org`.

	AGT	PER	ORG	ATT	JDG	EXP	WTN	PTY	AMC	JUR	LEG	CRT
full agreement	.74	.53	.59	.63	.80	.00	.00	.81	.63	.00	.48	.71
partial agreement	.87	.64	.74	.67	.84	.00	.00	.90	.71	.89	.48	.81

Table 3. The inter-annotator agreement for each of the agent mention types showing Agent (AGT), Person (PER), Organization (ORG), Attorney (ATT), Judge (JDG), Expert (EXP), Witness (WTN), Party (PTY), Amicus Curiae (AMC), Jury (JUR), Legislator (LEG), and Court (CRT).

2. *Maximum specificity* – the annotation should be done with the most specific appropriate type (e.g., in case the Agent, Organization, and Legislator types are all appropriate the Legislator type should be used).

For each type the guidelines provide a general definition as well as a couple of examples.[5]

Each decision was annotated by one of the annotators. A small subset (3 decisions) was annotated by both the annotators to measure inter-annotator agreement (see Table 3). We report the full as well as partial agreement. The full agreement is a ratio of the annotations that were created by the both users (i.e., they agree in type and the text span they cover) over all annotations. For partial agreement the annotations are considered to agree if they are of the same type and if they overlap by at least one character.

Table 3 shows that the agreement varies widely across the types. First, it should be noted that the type system is hierarchical. This means that any type also counts as the Agent. When computing the agreement for the Agent type we took into account all the 7004 annotated mentions (not just the 387 where the Agent type itself was marked). Something similar is true of the Person and the Organization types. The .00 agreement for the Expert and the Witness type is due to data sparsity. The agreement was measured on the IP documents. Table 4 shows that these two types were rare on these texts. The .00 full agreement (versus .89 partial agreement) for the Jury type is a systematic error of one of the annotators. The articles ("a", "an", "the") were supposed to be included in the annotations but the annotator failed to do so for the Jury type. As could be expected this error manifests in full agreement but it has no effect on partial agreement.

Table 4 provides detailed statistics of the created annotations. A rather small number of decisions (10) may suggest a relatively small size of the data set. As shown in Table 2 some of the decisions are very long. The total number of annotations (7004) clearly shows that the data set is sufficient for far more than toy experiments. The data set is publicly available.[6]

5. Experiments

5.1. Experimental Designs

We conducted three experiments to test the three hypotheses in this paper. In the *same domain experiment* we assessed the possiblity of detecting the agent mentions (types from Table 1) automatically (hypothesis 1). The goal of this experiment was to determine how well could a sequence labeling model (CRF) separate the signal from the noise for

[5]For example, the definition for the Attorney type is the following: "The Attorney type is reserved for mentions of agents that are known to be attorneys. These usually represent one of the parties or other participants of the proceedings (e.g., amicus curiae)."

[6]Hosted at `luima.org`.

	AGT	PER	ORG	ATT	JDG	EXP	WTN	PTY	AMC	JUR	LEG	CRT
cyber-crime												
# of seq	146	612	236	72	96	14	195	1352	0	82	17	334
# of seq / doc	29.2	122.4	47.2	14.4	19.2	2.8	39.0	270.4	0.0	16.4	3.4	66.8
intellectual-property												
# of seq	241	661	433	76	115	37	34	1668	35	81	16	451
# of seq / doc	48.2	132.2	86.6	15.2	23.0	7.4	6.8	333.6	7.0	16.2	3.2	90.2
total												
# of seq	387	1273	669	148	211	51	229	3020	35	163	33	785
# of seq / doc	38.7	127.3	66.9	14.8	21.1	5.1	22.9	302.0	5.5	16.3	3.3	78.5

Table 4. The summary statistics of the manually annotated agent mentions shows counts for Agent (AGT), Person (PER), Organization (ORG), Attorney (ATT), Judge (JDG), Expert (EXP), Witness (WTN), Party (PTY), Amicus Curiae (AMC), Jury (JUR), Legislator (LEG), and Court (CRT).

the purpose of recognizing the agent mentions. For this experiment the decisions were divided according to the domain from which they came.

In the *different domain experiment* we applied models trained on one area of law to the texts from the other domain. For example, we trained models on a training set of cyber-crime decisions and we evaluated them on on an intellectual property test set. The aim of this experiment was to confirm that the models' performance deteriorates when they are applied to decisions from a different domain (hypothesis 2). If so, it would suggest that the task is domain dependent.

In the *combined domains experiment* we used labeling models trained on one area of law to inform models trained for a different area. For example, predictions of a model trained on the cyber-crime data set were used as features for a model trained on the intellectual property data set. The goal of this experiment was to find out if a model improves when knowledge of another model trained for a different domain is taken into account (hypothesis 3).

In all the three experiments we train a separate CRF model for each agent mention type. Although this is certainly suboptimal, we use the same training strategy and features for all the models. It may be the case that different types (such as the Court or the Attorney) could benefit from a custom-tailored model and contextual features. We reserve fine-tuning of the individual models for future work. A CRF is a random field model that is globally conditioned on an observation sequence O. The states of the model correspond to event labels E. We use a first-order CRF in our experiments (observation O_i is associated with E_i). We use the CRFsuite[7] implementation of CRF. [11,12]

The texts were first tokenized. Each of the tokens is then a data point in a sequence a model operates on and it is represented by a small set of relatively simple low-level textual features. As labels we use the annotation types projected into the BILOU[8] scheme. The features include a token in lowercase, token's signature (a digit maps to "D", lowercase character maps to "c", uppercase to "C"), the token's length, its position within document, whether it is upper case, lowercase, titled, a digit or whitespace. For each token similar features from the three preceding and the three following tokens are included.

[7] `www.chokkan.org/software/crfsuite/`

[8] B: beginning of sequence, I: inside sequence, L: last in sequence, O: outside of sequence, U: unit-length sequence

5.2. Evaluation

To measure performance we use traditional IR metrics—precision (P), recall (R), and F_1-measure (F_1).

$$P = \frac{|Pred \cap Gold|}{|Pred|} \quad R = \frac{|Pred \cap Gold|}{|Gold|} \quad F = \frac{2 * P * R}{P + R}$$

Pred is the set of predicted annotations and *Gold* is the set of manually created annotations. In order to determine equality of annotations we used the same two approaches as when computing the inter-annotator agreement—the full (exact) match and the partial (overlap) match.

In the *same domain experiment* we used the leave one out cross-validation on the level of documents. This means that we have conducted the experiment for each of the documents. In a single round one document was a test set and the remaining documents from the same domain were included in the training set. For each type of agent we trained a separate CRF model on the training set. The model was then evaluated on the test set. The point was to see how successful the models are in detecting the agent mentions as compared to the performance of human experts.

For the *different domain experiment* a similar method was used. Again, the experiment was conducted multiple times—once for each document. Instead of using the remaining documents from the same domain as the training set, the documents from the other domain were used. The idea is to compare the performance of these models to the performance of the models trained on the same domains (the preceding experiment).

In the *combined domains experiment* the data from both domains were pooled together. Again, for each document there was a separate round. The point is to compare the performance of these models to that of the models trained on the same domains as well as on the different domains (the two preceding experiments) when applied alone. Our intuition was that at least some knowledge learned in other domain could be transferable.

5.3. Results

Table 5 summarizes the results of the three experiments described in Subsection 5.1. The evaluation metrics are explained in Subsection 5.2. The performance of the models differs considerably across the types but it correlates well across the experiments. That is, if the models trained to detect, say, the Jury type perform well in one of the experiments they perform similarly well in the other two experiments.

Because the type system is hierarchical we took into account all the predicted mentions when computing the metrics for the Agent type (i.e., notwithstanding its type any mention is also an agent). This is also true for the Person and the Organization types. All the other types are at the bottom level. Therefore only those mentions specifically marked with the respective type were considered when assessing the respective models.

The Jury and the Court models are very promising. The Agent, the Organization, the Attorney, the Judge, and the Party models have reasonable performance as well. The performance of the models for the Person and the Legislator types is lower but the models obviously are able to pick some signal. The models for the remaining types perform poorly. In case of the Expert and the Witness types, data sparsity could be the cause.

The models created in the *different domain experiment* tend to have the lowest performance (the middle block of Table 5). This is especially true for the Agent, the Person,

		AGT	PER	ORG	ATT	JDG	EXP	WTN	PTY	AMC	JUR	LEG	CRT
same domain													
exact	P	.74	.65	.79	.67	.47	.00	.56	.73	.17	.87	.50	.81
	R	.36	.17	.39	.25	.16	.00	.04	.36	.03	.56	.06	.69
	F_1	.48	.27	.52	.37	.23	.00	.08	.48	.05	.68	.11	.75
overlap	P	.83	.73	.85	.73	.72	.00	.61	.84	.50	.91	1.0	.87
	R	.40	.19	.42	.27	.24	.00	.05	.41	.09	.59	.12	.74
	F_1	.54	.31	.57	.39	.36	.00	.09	.55	.15	.72	.22	.80
different domain													
exact	P	.67	.48	.70	.59	.46	.00	.00	.63	.00	.85	.27	.80
	R	.28	.09	.39	.18	.20	.00	.00	.23	.00	.63	.09	.68
	F_1	.39	.16	.49	.27	.28	.00	.00	.33	.00	.73	.14	.73
overlap	P	.76	.58	.75	.64	.64	.00	.00	.74	.00	.90	.55	.85
	R	.31	.11	.42	.19	.28	.00	.00	.27	.00	.66	.18	.72
	F_1	.44	.19	.54	.29	.39	.00	.00	.39	.00	.77	.27	.78
combined domains													
exact	P	.70	.66	.73	.68	.52	.00	.52	.69	.22	.88	.45	.79
	R	.37	.23	.43	.35	.26	.00	.06	.34	.06	.69	.15	.72
	F_1	.48	.34	.54	.46	.34	.00	.11	.46	.09	.77	.23	.76
overlap	P	.79	.74	.78	.72	.73	.00	.52	.80	.44	.92	.64	.85
	R	.41	.25	.46	.37	.36	.00	.06	.39	.11	.72	.21	.78
	F_1	.54	.38	.58	.49	.48	.00	.11	.53	.18	.81	.32	.81

Table 5. The performance of the CRF models in automatic detection of agent mentions. The measures used are Precision (P), Recall (R), and F_1-measure (F_1). We assess the models trained to detect Agent (AGT), Person (PER), Organization (ORG), Attorney (ATT), Judge (JDG), Expert (EXP), Witness (WTN), Party (PTY), Amicus Curiae (AMC), Jury (JUR), Legislator (LEG), and Court (CRT).

and the Party types. The best performing models are those created in the *combined domains experiment*. All the models perform at least as well as those that were generated in the *same domain experiment*. The models for the Person, the Attorney, the Judge, and the Jury perform significantly better.

6. Discussion and Future Work

The results summarized in Table 5 clearly show that simple CRF models using low-level textual features are capable of detecting different types of agent mentions automatically. In case of some types (Jury, Court) the performance appears to be sufficient for actual use. In case of some other types (Expert, Witness, Legislator) the performance is clearly too low to produce useful results. For the remaining types it is not clear if the results would have the potential to be useful in practice. This may also depend on the intended application (attribution resolution, summarization).

The performance of the models generated during the *same domain experiment* (top part of Table 5) is better than the performance of the models trained in the *different domain experiment* (middle part of Table 5). This suggests that for each domain there may be certain agent mentions that are rare or non-existent in other domains. In cyber crime one of the prosecuting parties was often mentioned as "the government." This rarely happens in the IP disputes where two private parties are usually involved.

The models created during the *combined domains experiment* (bottom part of Table 5) generally outperformed the models created during both, the *same domain experiment* as well as the *different domain experiment*. This shows that certain patterns in mentioning agents transfer across domains. The Court type mentions appear to transfer very well since even the models trained on the different domain were capable of retaining good performance (e.g., "we" is universally being used to mention the deciding majority).

It is worth emphasizing that the models trained in our experiments are quite simplistic, especially in terms of features they use. While examining the errors it became very clear that simple textual features do not provide sufficient information to detect certain mentions and to distinguish among the different types. One could easily see how using additional resources could lead to dramatic improvements. Take the Amicus Curiae type as an example. The models struggled to distinguish the mentions of this type from mentions of other types, especially the Party and the Organization type. Yet the amici are almost always listed in the header of the decision in a manner that could often allow detection through simple regular expression matching. It is quite likely that detection of the amici in the header and using the detected tokens as contextual features could raise the performance of our models from very bad to excellent.

There are multiple aspects of this work that we would like to address (or see addressed) in future. For some of the mention types (Expert, Witness) we encountered the data sparsity problem. This issue could be affecting other types, too, even though it does not manifest that clearly. It would make sense to enrich the data set with additional documents (perhaps from other areas of law). An interesting option would be to include annotated documents from courts outside the U.S. (e.g., the EU's Court of Justice).

We have defined the limited type hierarchy that includes only the most basic types of agents that are regularly mentioned in decisions (see Table 1). These are by no means all the types that would be of interest for automatic detection. Some of the types that are already included could be further differentiated into subcategories (e.g., Party to Plaintiff, Defendant, Apellant). Thus extending the type hierarchy and annotating the corpus with the new (extended) types would be another way to continue in this work.

The models that we used are fairly simple, especially in terms of the low-level textual features they operate on. Above we have discussed how using more advanced features could lead to considerable improvements. Although, CRF is a decent model for this task some more recent sequence labeling models (e.g., long short-term memory networks) are likely to perform even better provided there is enough data to train them.

Assuming we are able to detect the agent mentions with sufficient accuracy, coreference resolution is a traditional task in natural language processing. The goal in coreference resolution is to determine which words or phrases refer to the same object. In the context of agent mentions this would mean finding out which mentions denote the same agent (e.g., mentions such as "we", "our", "the majority", "this court" could all denote the same agent in a decision).

The ultimate goal is to apply this work in practice. One such application could be automatic attribution resolution. It would be of immense value for a system to determine if a certain interpretation of a legal rule is advanced by the deciding majority, a dissenting judge, or one of the parties. Successful attribution resolution would greatly improve legal IR, argumentation mining, or automatic summarization of legal documents.

7. Conclusions

In this paper we examined the possibility of automatically detecting agent mentions in case law analysis. We have shown that: (i) with varying degree of accuracy it is possible to detect the mentions of different agent types automatically; (ii) the task is domain dependent in a sense that prediction models trained on one area of law do not perform as

well for a different area; and (iii) there is relatedness between domains allowing the use of data from one area of law to improve performance of a model intended for another area. It is our hope that this work will stimulate further research in detecting agent mentions in legal texts. For this reason we release the data set that was created to facilitate the experiments described in this paper. We leave plenty of space for further improvements.

Aknowledgements

This work was supported in part by the National Institute of Justice Graduate Student Fellowship (Fellow: Jaromir Savelka) Award # 2016-R2-CX-0010, "Recommendation System for Statutory Interpretation in Cybercrime."

References

[1] Al-Kofahi, Khalid, Brian Grom, and Peter Jackson. "Anaphora resolution in the extraction of treatment history language from court opinions by partial parsing." *Proceedings of the 7th international conference on Artificial intelligence and law*. ACM, 1999.

[2] Ashley, Kevin D. *Artificial Intelligence and Legal Analytics*. Cambridge University Press, 2017.

[3] Ashley, Kevin D. *Modeling legal arguments: Reasoning with cases and hypotheticals*. MIT press, 1991.

[4] Ashley, Kevin D., and Vern R. Walker. "From Information Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study." *JURIX*. 2013.

[5] Ashley, Kevin D., and Vern R. Walker. "Toward constructing evidence-based legal arguments using legal decision documents and machine learning." *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*. ACM, 2013.

[6] Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. "Named entity recognition and resolution in legal text." Semantic Processing of Legal Texts. Springer Berlin Heidelberg, 2010. 27–43.

[7] Faiz, Syeed Ibn, and Robert Mercer. "Extracting higher order relations from biomedical text." *Proceedings of the First Workshop on Argumentation Mining*. 2014.

[8] Francesconi, Enrico. "An Approach to Legal Rules Modelling and Automatic Learning." *JURIX*. 2009.

[9] Grabmair, Matthias, Ashley, K. D., Chen, R., Sureshkumar, P., Wang, C., Nyberg, E., and Walker, V. R. "Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools." *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM, 2015.

[10] Grabmair, M., Ashley, K. D., Hwa, R., and Sweeney, P. M. "Toward Extracting Information from Public Health Statutes using Text Classification Machine Learning." *JURIX*. 2011.

[11] John Lafferty, Andrew McCallum, Fernando Pereira, and others. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning*, ICML, Vol. 1. 282–289.

[12] Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields. (2007).

[13] Peters, Wim, and Adam Z. Wyner. "Legal Text Interpretation: Identifying Hohfeldian Relations from Text." *LREC*. 2016.

[14] Quaresma, Paulo, and Teresa Goncalves. "Using linguistic information and machine learning techniques to identify entities from juridical documents." *Semantic Processing of Legal Texts*. Springer Berlin Heidelberg, 2010. 44–59.

[15] Savelka, Jaromir, Matthias Grabmair, and Kevin D. Ashley. "Mining Information from Statutory Texts in Multi-Jurisdictional Settings." *JURIX*. 2014.

[16] Walker, Vern R. "The Need for Annotated Corpora from Legal Documents, and for (Human) Protocols for Creating Them: The Attribution Problem." *Dagstuhl Seminar on Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments*, 2016.

[17] Walker, Vern R., Parisa Bagheri, and Andrew J. Lauria. "Argumentation Mining from Judicial Decisions: The Attribution Problem and the Need for Legal Discourse Models." 2015.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-49

Temporalised Belief Revision in the Law

Luciano H. TAMARGO [a,1], Diego C. MARTINEZ [a] Antonino ROTOLO [b]
Guido GOVERNATORI [c]
[a] *UNS–CONICET, Universidad Nacional del Sur, Argentina*
[b] *CIRSFID, University of Bologna, Italy*
[c] *Data61, CSIRO, Australia*

Abstract. This paper presents a belief revision operator for legal systems that considers time intervals. This model relates techniques about belief revision formalisms and time intervals with temporalised rules for legal systems. Our goal is to formalise a temporalised belief base and corresponding timed derivation, together with a proper revision operator. This operator may remove rules when needed or adapt intervals of time when contradictory norms are added in the system.

Keywords. Norm Change, Belief Revision, Temporal Reasoning

1. Introduction and Motivation

One peculiar feature of the law is that it necessarily takes the form of a dynamic normative system [22,21]. Despite the importance of norm-change mechanisms, the logical investigation of legal dynamics was for long time underdeveloped. However, research is rapidly evolving and recent contributions exist.

In the eighties a pioneering research effort was devoted by Alchourrón, Gärdenfors and Makinson [4] to develop a logical model (AGM) for also modeling norm change. As is well-known, the AGM framework distinguishes three types of change operation over theories. Contraction is an operation that removes a specified sentence ϕ from a given theory Γ (a logically closed set of sentences) in such a way that Γ is set aside in favor of another theory Γ^{-}_{ϕ} which is a subset of Γ not containing ϕ. Expansion operation adds a given sentence ϕ to Γ so that the resulting theory Γ^{+}_{ϕ} is the smallest logically closed set that contains both Γ and ϕ. Revision operation adds ϕ to Γ but it is ensured that the resulting theory Γ^{*}_{ϕ} be consistent. Alchourrón, Gärdenfors and Makinson argued that, when Γ is a code of legal norms, contraction corresponds to norm derogation (norm removal) and revision to norm amendment. AGM framework has the advantage of being very abstract, as it works with theories consisting of simple logical assertions. For this reason, it can capture basic aspects of the dynamics of legal systems, such as the change obligations and permissions [7,14].

Some research has been carried out to reframe AGM ideas within richer rule-based logical systems [24,23]. However, also these attempts suffer from some drawbacks of

[1]Corresponding Author: Luciano H. Tamargo, Institute for Computer Science and Engineering (UNS–CONICET), Department of Computer Science and Engineering, Universidad Nacional del Sur, Argentina; E-mail: lt@cs.uns.edu.ar

standard AGM, among them the fact that the proposed frameworks fail to handle the temporal aspects of norm change: indeed, legal norms are qualified by temporal properties, such as the time when the norm comes into existence and belongs to the legal system, the time when the norm is in force, the time when the norm produces legal effects, and the time when the normative effects hold. Since all these properties can be relevant when legal systems change, [14] argues that failing to consider the temporal aspects of legal dynamics poses a serious limit to correctly model norm change in the law.

Unlike rich but complex frameworks such as the one of [14], this paper claims that belief revision techniques—which are based on an abstract and elegant machinery—can be reconciled with need to consider several temporal patterns of legal reasoning. In this work we are thus interested in the formalisation of a belief revision operator applied to an epistemic model that considers rules and time. We enrich a simple logic language with an interval-based model of time, to represent validity and effectiveness of a norm. The revision operator may remove rules when needed or adapt intervals of time when newer, contradictory norms are introduced in the system.

The layout of the paper is as follows. Section 2 shows an example to motivate the main ideas of our proposal. Section 3 proposes the notions of temporalised belief base and temporalised derivation. Section 4 presents a norm revision operator based on temporalised belief base. Section 5 reports on related work. Some conclusions end the paper.

2. Motivating Example

Let us first of all present a concrete example that will serve to motivate the main ideas of our proposal.

EXAMPLE 1. *Consider the following pieces of information regarding a legislative attempt to ease tax pressure for people that have been unemployed.*

- *A citizen was unemployed from 1980 to 1985.*
- *If unemployed from 1980 to 1983, then a tax exemption applies from 1984 to 1986, in order to increase individual savings.*
- *New authorities in government revoke tax exemption for years 1985 and 1986.*
- *Tax exemption reinstated for the year 1985 due to agreements with labor unions.*

However, later on the legislators approved a new normative establishing that finally there is no tax-exemption for all citizens for the years 1985 and 1986.

The previous situation seems to establish that, at the end, a tax exemption applies only for year 1985 for a while, before being revoked.

3. Legal System as Temporalised Belief Base

The problem of representing temporal knowledge and temporal reasoning arises in many disciplines, including Artificial Intelligence. A usual way to do this is to determine a *primitive* to represent time, and its corresponding *metric relations*. There are in the literature two traditional approaches to reasoning with and about time: a point based approach, as in [14], and an interval based approach as in [5,9]. In the first case, the emphasis is put on *instants* of time (e.g., timestamps) and a relation of precedence among them. In

the second case, time is represented as continuous sets of instants in which something relevant occurs. These intervals are identified by the starting and ending instants of time.

In this work, time intervals (like in [6,9]) will be considered. This design decision has been taken because it simplifies the construction of an revision operator which will be introduced below. That is, following the semantics of the temporalised rules proposed in [14] and explained in Section 3 (an adapted version), the revision operator in many cases only consists in modifying the intervals to maintain the consistency.

Besides, different temporal dimensions will be taken into account. That is, as it is mentioned in [14], in a normative system, norms have different temporal dimensions: the time of **validity** of a norm (when the norm enters in the normative system) and the time of **effectiveness** (when the norm can produce legal effects). Thus, if one wants to model norm modifications, then normative systems must be modelled by more complicated structures. In particular, a normative system is not just the set of norms valid in it, but it should also consider the normative systems where the norms are effective.

3.1. Preliminaries and Notation

We will adopt a propositional language $\mathbb{L}$ with a complete set of boolean connectives: $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$. Each formula in $\mathbb{L}$ will be denoted by lowercase Greek characters: $\alpha, \beta, \delta, \ldots, \omega$. We will say that α is the complement of $\neg\alpha$ and vice versa. The characters σ will be reserved to represent cut function for a change operator. We also use a consequence operator, denoted $Cn(\cdot)$, that takes sets of sentences in $\mathbb{L}$ and produces new sets of sentences. This operator $Cn(\cdot)$ satisfies *inclusion* ($A \subseteq Cn(A)$), *idempotence* ($Cn(A) = Cn(Cn(A))$), and *monotony* (if $A \subseteq B$ then $Cn(A) \subseteq Cn(B)$). We will assume that the consequence operator includes classical consequences and verifies the standard properties of *supraclassicality* (if α can be derived from A by deduction in classical logic, then $\alpha \in Cn(A)$), *deduction* ($\beta \in Cn(A \cup \{\alpha\})$ if and only if $(\alpha \rightarrow \beta) \in Cn(A)$) and *compactness* (if $\alpha \in Cn(A)$ then $\alpha \in Cn(A')$ for some finite subset A' of A). In general, we will write $\alpha \in Cn(A)$ as $A \vdash \alpha$. Note that the AGM model [4] represents epistemic states by means of belief sets, that is, sets of sentences closed under logical consequence. Other models use belief bases; i.e., arbitrary sets of sentences [10,18]. Our epistemic model is based on an adapted version of belief bases which have additional information (time intervals). The use of belief bases makes the representation of the legal system state more natural and computationally tractable. That is, following [20] (page 24), we consider that legal systems sentences could be represented by a limited number of sentences that correspond to the explicit beliefs of the legal system.

3.2. Time Interval

We will consider a universal finite *set of time labels* $\mathbb{T} = \{t_1, \ldots, t_n\}$ strictly ordered; each time label will represent an unique time instant. Simplifying the notation, we assume that $t_i - 1$ is the immediately previous instant to the instant t_i and $t_i + 1$ is the immediately posterior instant to the instant t_i.

Like in [16] we propose temporalised literals, however, we use intervals. We will consider an interval like finite ordered sequence of time labels $t_i, \ldots, t_j$ where i, j are natural numbers ($i \leq j$) and $t_i, \ldots t_j \in \mathbb{T}$ denoting instances of time or *timepoints*. Thus we have expressions of the type $\alpha^{interval}$, where *interval* can be as follow:

- $[t_i,t_i]$: meaning that α holds at time t_i. Following [14] α is transient (holding at precisely one instant of time). For simplicity $[t_i,t_i] = [t_i]$.
- $[t_i,\infty]$: meaning that α holds from t_i. Following [14] α is persistent from t_i.
- $[t_i,t_j]$: meaning that α holds from time t_i to t_j with $t_i < t_j$.

Then we will consider a set of time intervals $\mathbb{I}$ which contains intervals as those described previously. Thus, for simplicity, we can have expressions like α^J where $J \in \mathbb{I}$. Intervals in $\mathbb{I}$ will be denoted by uppercase Latin characters: $A,B,C,\ldots,Z$. Two intervals may not be disjoint, as defined next.

Definition 1. *Contained interval. Let $R,S \in \mathbb{I}$ be two intervals. We say that R is contained in S, denoted $R \subseteq S$ if and only if for all $t_i \in R$ it holds that $t_i \in S$.*

Definition 2. *Overlapped interval. Let $R,S \in \mathbb{I}$ be two intervals. We say that R and S are overlapped, denoted $R \approx S$ if and only if there exists $t_i \in R$ such that $t_i \in S$.*

EXAMPLE 2. *Let $R,S,V \in \mathbb{I}$ where $R = [t_3,t_7]$, $S = [t_4,t_6]$ and $V = [t_5,t_9]$ with $t_3,t_4,t_5,t_6,t_7,t_9 \in \mathbb{T}$. Then $S \subseteq R$, $R \approx V$ and $S \approx V$.*

3.3. Temporalised Belief Base

As rules are part of the knowledge, they are subject of temporal validity too: the *time of force* of a rule, i.e., the time when a rule can be used to derive a conclusion given a set of premises. In this perspective we can have expressions like

$$(\alpha^{[t_a,t_b]} \rightarrow \beta^{[t_c,t_d]})^{[t_e,t_f]}$$

meaning that the rule is in force from timepoint t_e to t_f, or in other words, we can use the rule to derive the conclusion at time from time t_e to t_f. The full semantics of this expression is that from time t_e to t_f we can derive that β holds from time t_c to t_d if we can prove that α holds from time t_a to t_b. But now we are doing a derivation from time t_e to t_f, so the conclusion $\beta^{[t_c,t_d]}$ is derived from time t_e to t_f and the premise $\alpha^{[t_a,t_b]}$ must be derived from time t_e to t_f as well. In the same way a conclusion can persist, this applies as well to rules and then to derivations.

Thus, it is possible to define *temporalised belief base* which will contain temporalised literal and temporalised rules (see Example 3). This base represents a legal system in which each temporalised sentence defines a norm whose time interval determines the validity and effectiveness time.

EXAMPLE 3. *A legal system can be represented by the temporalised belief base* $\mathbb{K} = \{\alpha^{[t_1,t_3]}, \alpha^{[t_4]}, (\alpha^{[t_1,t_4]} \rightarrow \beta^{[t_4,t_6]})^{[t_4,t_6]}, \beta^{[t_5,t_6]}, \beta^{[t_6,t_8]}, \beta^{[t_{10}]}, \delta^{[t_{11}]}, (\delta^{[t_{11}]} \rightarrow \beta^{[t_{15},t_{20}]})^{[t_5,t_6]}, \omega^{[t_2,t_8]}, (\omega^{[t_4]} \rightarrow \beta^{[t_6,\infty]})^{[t_{21},t_{22}]}, \varepsilon^{[t_1,\infty]}\}$.

This type of belief base representation implies that a sentence can appear more than once in a temporalised belief base; but from the point of view of the temporalised sentences stored in the temporalised belief base there is no redundancy because each temporalised sentence has different time interval. For instance, consider Example 3, α appears two times, but with different intervals. In this case, we will say that α is **intermittent** and it means that α is held from t_1 to t_3 and it is held in the instant t_4. Besides, if the intervals of a sentence are overlapped ($\beta^{[t_5,t_6]}$, $\beta^{[t_6,t_8]}$ in Example 3), despite that the time interval of the sentence intuitively be only one ($[t_5,t_8]$), we decided to maintain all versions because will be more suitable when we will model the dynamics of the legal system.

3.4. Temporalised Derivation

Note that a norm can explicitly be in a temporalised belief base, $\alpha^{[t_5]} \in \mathbb{K}$ in Example 3. However, a norm can implicitly be represented in a temporal belief base if some conditions are held. For instance, in Example 3, β is implicitly represented with $\omega^{[t_2,t_8]}$, $(\omega^{[t_4]} \rightarrow \beta^{[t_6,\infty]})^{[t_{21},t_{22}]}$ due to the antecedent of the rule is held in t_4 by the temporalised sentence $\omega^{[t_2,t_8]}$. Next, temporalised derivation for a sentence are defined to capture this notion. To do this, first we define a temporalised derivation in a time instant and then we give a definition of temporalised derivation in time interval. The last mentioned is based on the previous.

Definition 3. *Temporalised derivation in a time instant. Let $\mathbb{K}$ be a set of temporalised sentences and $\alpha^{[t_i]}$ be a temporalised sentence. We say that $\alpha^{[t_i]}$ is derived from $\mathbb{K}$ (denoted $\mathbb{K} \vdash^t \alpha^{[t_i]}$) if and only if:*

- *$\alpha^J \in \mathbb{K}$ and $t_i \in J$, or*
- *$(\beta^H \rightarrow \alpha^P)^Q \in \mathbb{K}$ and $t_i \in P$ and $\mathbb{K} \vdash^t \beta^{[t_j]}$ for all $t_j \in H$.*

Definition 4. *Temporalised derivation in a time interval. Let $\mathbb{K}$ be a set of temporalised sentences and $\alpha^{[t_i,t_j]}$ be a temporalised sentence. We say that $\alpha^{[t_i,t_j]}$ is derived from $\mathbb{K}$ (denoted $\mathbb{K} \vdash^t \alpha^{[t_i,t_j]}$) if and only if $\mathbb{K} \vdash^t \alpha^{[t_p]}$ for all $t_p \in [t_i,t_j]$.*

To compute the temporalised derivation of a sentence checking each instant of the intervals benefits us in special cases where implicit sentences need temporalised sentences with overlapped intervals as antecedents. To determine the time interval of the implicitly derived temporal sentence, the temporal consequence will be defined below.

Definition 5. *Temporalised consequence. Let $\mathbb{K}$ be a set of temporalised sentences and $\alpha^{[t_i,t_j]}$ be a temporalised sentence. We say that $\alpha^{[t_i,t_j]}$ is a temporalised consequence of $\mathbb{K}$ ($\alpha^{[t_i,t_j]} \in Cn^t(\mathbb{K})$) if and only if $\mathbb{K} \vdash^t \alpha^{[t_i,t_j]}$.*

EXAMPLE 4. *Consider again the temporalised belief base of Example 3. Then, $\mathbb{K} \vdash^t \beta^{[t_4,\infty]}$, that is, $\beta^{[t_4,\infty]} \in Cn^t(\mathbb{K})$; and $\mathbb{K} \vdash^t \alpha^{[t_1,t_4]}$, that is, $\alpha^{[t_1,t_4]} \in Cn^t(\mathbb{K})$.*

Note that the underlying semantics of this type of derivation (legal system) differs from that in propositional logic when we want to represent the knowledge [9]. Note that, following Definition 4, the **interval of an implicitly derived sentence** will be the interval of the consequent of the rule that derives the conclusion of the proof. For instance, suppose that $\mathbb{K} = \{\gamma^{[t_2,t_5]}, (\gamma^{[t_3,t_4]} \rightarrow \varepsilon^{[t_6,t_9]})^{[t_1,\infty]}\}$ then the time interval of ε is $[t_6,t_9]$.

In this proposal, a ***contradiction*** arise when two complementary sentences can be derived with time intervals overlapped. For instance, suppose $\mathbb{K} = \{\alpha^{[t_2,t_9]}, \neg\alpha^{[t_1,t_3]}\}$, in this case there exist a contradiction. However, consider $\mathbb{K} = \{\alpha^{[t_5]}, \neg\alpha^{[t_1,t_3]}\}$, in this case, we will say that $\mathbb{K}$ does not have contradictions. Moreover, we will say that a temporalised belief base is ***temporally consistent*** if the base does not have contradictions. The temporalised belief base of Example 3 is temporally consistent.

4. Legal Belief Revision

A legal system should be temporally consistent, i.e., it cannot contain contradictory norms at any time. Hence, we propose a **norm prioritised revision operator** that allows to consistently add a temporalised sentence $\alpha^{[t_i,t_j]}$ to a consistent legal system $\mathbb{K}$.

This special revision operator is inspired in the rule semantics explained above in Section 3 (an adapted version from that proposed in [14]). Thus, following the concept of consistency proposed in Section 3, the revision operator may remove temporalised sentences or, in some cases, may only modify the intervals to maintain consistency.

To incorporate a norm $\neg\beta^J$ into a legal system, it is necessary to consider all possible contradictions that may arise if the norm is added without checking for consistency. For this, it is necessary to compute all proofs of β considering only those temporalised sentences β^P whose effectiveness time is overlapped with the time interval J, that is, $J \approx P$. Note that, computing all minimal proofs of a temporal sentence considering only those which time interval is overlapped with the time interval of the input sentence, is an optimized version. Next, a set of minimal proof for a sentence will be defined.

Definition 6. *Let $\mathbb{K}$ be a temporalised belief base and α^J a temporalised sentence. Then, $\mathbb{H}$ is a minimal proof of α^J if and only if*

1. $\mathbb{H} \subseteq \mathbb{K}$,
2. $\alpha^P \in Cn^t(\mathbb{H})$ *with* $J \approx P$, *and*
3. *if* $\mathbb{H}' \subset \mathbb{H}$, *then* $\alpha^P \notin Cn^t(\mathbb{H}')$ *with* $J \approx P$.

Given a temporalised sentence α^J, the function $\Pi(\alpha^J, \mathbb{K})$ returns the set of all the minimal proofs for α^J from $\mathbb{K}$.

REMARK 1. *Each set of $\Pi(\alpha^J, \mathbb{K})$ derives α in at least one time instant of J.*

EXAMPLE 5. *Consider the temporalised belief base of Example 3. Then $\Pi(\beta^{[t_5,t_6]}, \mathbb{K})$ = $\{\mathbb{H}_1, \mathbb{H}_2, \mathbb{H}_3, \mathbb{H}_4\}$ where:*

- $\mathbb{H}_1 = \{\alpha^{[t_1,t_3]}, \alpha^{[t_4]}, (\alpha^{[t_1,t_4]} \rightarrow \beta^{[t_4,t_6]})^{[t_4,t_6]}\}$,
- $\mathbb{H}_2 = \{\beta^{[t_5,t_6]}\}$,
- $\mathbb{H}_3 = \{\beta^{[t_6,t_8]}\}$,
- $\mathbb{H}_4 = \{\omega^{[t_2,t_8]}, (\omega^{[t_4]} \rightarrow \beta^{[t_6,\infty]})^{[t_{21},t_{22}]}\}$

Note that $\mathbb{H}_1$ is minimal due to α should be derived from t_1 to t_4 to use the rule $(\alpha^{[t_1,t_4]} \rightarrow \beta^{[t_4,t_6]})^{[t_4,t_6]}$ hence, $\alpha^{[t_1,t_3]}$ and $\alpha^{[t_4]}$ should be in $\mathbb{H}_1$.

Our operator is based on a selection of sentences in the knowledge base that are relevant to derive the sentence to be retracted or modified. In order to perform a revision, following kernel contractions [19], this approach uses *incision functions*, which select from the minimal subsets entailing the piece of information to be revoked or modified. We adapt this notion of incision function proposed in [19] to our epistemic model. An incision function only selects sentences that can be relevant for α and at least one element from each $\Pi(\alpha^J, \mathbb{K})$, as follows.

Definition 7. *Incision function. Let $\mathbb{K}$ be a temporalised belief base, an incision function σ for $\mathbb{K}$ is a function such that for all $\alpha^J \in Cn^t(\mathbb{K})$:*

- $\sigma(\Pi(\alpha^J, \mathbb{K})) \subseteq \bigcup(\Pi(\alpha^J, \mathbb{K}))$.
- *For each* $\mathbb{H} \in \Pi(\alpha^J, \mathbb{K})$, $\mathbb{H} \cap \sigma(\Pi(\alpha^J, \mathbb{K})) \neq \emptyset$.

In Hansson's work it is not specified how the incision function selects the sentences that will be discarded of each minimal proof. In our approach, this will be solved by considering those sentences that can produce legal effects in favour of a possible contradiction with the new norm. Thus, if the new norm is $\neg\beta^J$ then the incision function will select the temporalised sentences β^P or $(\alpha^Q \to \beta^F)^H$ of each $\Pi(\beta^J, \mathbb{K})$.

Definition 8. *Search consequence function. Sc:* $\mathbb{L} \times \mathbb{K} \mapsto \mathbb{K}$*, is a function such that for a given sentence* α *and a given temporalised base* $\mathbb{K}$ *with* $\mathbb{H} \subseteq \mathbb{K}$*,*

$$Sc(\alpha, \mathbb{H}) = \{\alpha^J : \alpha^J \in \mathbb{H}\} \cup \{(\beta^P \to \alpha^Q)^R : (\beta^P \to \alpha^Q)^R \in \mathbb{H} \text{ and } \beta \in \mathbb{L}\}$$

Definition 9. *Consequence incision function. Given a set of minimal proofs* $\Pi(\alpha^J, \mathbb{K})$*,* σ^c *is a consequence incision function if it is a incision function for* $\mathbb{K}$ *such that*

$$\sigma^c(\alpha^J, \mathbb{K}) = \bigcup_{\mathbb{H} \in \Pi(\alpha^J, \mathbb{K})} Sc(\alpha, \mathbb{H})$$

EXAMPLE 6. *Consider Examples 3 and 5. Then,* $Sc(\beta, \mathbb{H}_1) = \{(\alpha^{[t_1,t_4]} \to \beta^{[t_4,t_6]})^{[t_4,t_6]}\}$*,* $Sc(\beta, \mathbb{H}_2) = \{\beta^{[t_5,t_6]}\}$*,* $Sc(\beta, \mathbb{H}_3) = \{\beta^{[t_6,t_8]}\}$*, and* $Sc(\beta, \mathbb{H}_4) = \{(\omega^{[t_4]} \to \beta^{[t_6,\infty]})^{[t_{21},t_{22}]}\}$*. Thus,* $\sigma^c(\beta^{[t_5,t_6]}, \mathbb{K}) = \bigcup_{\mathbb{H} \in \Pi(\beta^{[t_5,t_6]}, \mathbb{K})} Sc(\beta, \mathbb{H}) = \{(\alpha^{[t_1,t_4]} \to \beta^{[t_4,t_6]})^{[t_4,t_6]}, \beta^{[t_5,t_6]}, \beta^{[t_6,t_8]}, (\omega^{[t_4]} \to \beta^{[t_6,\infty]})^{[t_{21},t_{22}]}\}$

As mentioned before, the revision operator may remove temporalised sentences or, in some cases, may modify the intervals to maintain consistency. Next, a temporal projection will be defined based on a given time interval. The idea here is, given a temporalised belief base $\mathbb{K}$ and given a time interval $[t_i, t_j]$, to return a temporalised belief base $\mathbb{K}'$ containing those sentences from $\mathbb{K}$ whose time intervals be out of $[t_i, t_j]$.

Definition 10. *Excluding temporal projection. Let* $\mathbb{K}$ *be a temporalised belief base and let* $[t_i, t_j]$ *be a time interval where* $t_i, t_j \in \mathbb{T}$*. A excluding temporal projection of* $\mathbb{K}$ *from* t_i *to* t_j*, denoted* $\overset{out}{\mathbb{K}}{}^{t_i}_{t_j}$*, is a subset of* $\mathbb{K}$ *where for all* $\alpha^{[t_p,t_q]} \in \mathbb{K}$*,* $\overset{out}{\mathbb{K}}{}^{t_i}_{t_j}$ *will contain:*

- $\alpha^{[t_p,t_i-1]}$ *if* $t_p < t_i$*,* $t_q \geq t_i$ *and* $t_q \leq t_j$*.*
- $\alpha^{[t_j+1,t_q]}$ *if* $t_p \geq t_i$*,* $t_q > t_j$ *and* $t_p \leq t_j$*.*
- $\alpha^{[t_p,t_i-1]}$ *and* $\alpha^{[t_j+1,t_q]}$ *if* $t_p < t_i$*,* $t_q > t_j$*.*
- $\alpha^{[t_p,t_q]}$ *if* $t_q < t_i$ *or* $t_p > t_j$*.*

REMARK 2. *Note that the case in which* $t_p \geq t_i$ *and* $t_q \leq t_j$ *the temporal sentence it is not considered. In this case, this sentence is erased.*

EXAMPLE 7. *Consider Example 6 and suppose that S is a temporalised belief base and* $S = \sigma^c(\beta^{[t_5,t_6]}, \mathbb{K})$*. Then,* $\overset{out}{S}{}^{t_5}_{t_6} = \{(\alpha^{[t_1,t_4]} \to \beta^{[t_4]})^{[t_4,t_6]}, \beta^{[t_7,t_8]}, (\omega^{[t_4]} \to \beta^{[t_7,\infty]})^{[t_{21},t_{22}]}\}$.

Following the notion of excluding temporal projection (Definition 10) a norm prioritized revision operator can be defined. That is, an operator that allows to *consistently* add temporalised sentences in a temporalised belief base. If a contradiction arises, then the revision operator may remove temporalised sentences or modify the corresponding intervals in order to maintain consistency.

Definition 11. *Let $\mathbb{K}$ be a temporalised belief base and $\alpha^{[t_i,t_j]}$ be a temporalised sentence. The operator " $\otimes$", called prioritized revision operator, is defined as follow:*

$$\mathbb{K} \otimes \alpha^{[t_i,t_j]} = (\mathbb{K} \setminus S) \cup \overset{out}{S}{}^{t_i}_{t_j} \cup \{\alpha^{[t_i,t_j]}\}$$

where $S = \sigma^c(\neg\alpha^{[t_i,t_j]}, \mathbb{K})$

EXAMPLE 8. *Consider Example 3 and suppose that a new norm $\neg\beta^{[t_5,t_6]}$ it is wished to add. To do this, it is necessary to do $\mathbb{K} \otimes \neg\beta^{[t_5,t_6]}$. Consider Examples 5 and 6. Then, $\mathbb{K} \otimes \neg\beta^{[t_5,t_6]} = \{\alpha^{[t_1,t_3]}, \alpha^{[t_4]}, (\alpha^{[t_1,t_4]} \rightarrow \beta^{[t_4]})^{[t_4,t_6]}, \beta^{[t_7,t_8]}, \beta^{[t_{10}]}, \delta^{[t_{11}]}, (\delta^{[t_{11}]} \rightarrow \beta^{[t_{15},t_{20}]})^{[t_5,t_6]}, \omega^{[t_2,t_8]}, (\omega^{[t_4]} \rightarrow \beta^{[t_7,\infty]})^{[t_{21},t_{22}]}, \varepsilon^{[t_1,\infty]}, \neg\beta^{[t_5,t_6]}\}$. Note that, this new temporalised base is temporally consistent.*

The following example shows how our operator works in a particular situation when a legal system undergoes many changes and has rules that complement each other.

EXAMPLE 9. *Consider following temporalised belief base $\mathbb{K} = \{\beta^{[t_1,t_{10}]}, (\beta^{[t_1,t_5]} \rightarrow \alpha^{[t_1,t_5]})^{[t_1,\infty]}, (\beta^{[t_6,t_{10}]} \rightarrow \alpha^{[t_6,t_{10}]})^{[t_1,\infty]}, \delta^{[t_4]}\}$. Note that, $\mathbb{K} \vdash^t \alpha^{[t_1,t_{10}]}$ because $\mathbb{K} \vdash^t \alpha^{[t_i]}$ for all $t_i \in [t_1,t_{10}]$. Suppose that it is necessary to adopt $\neg\alpha^{[t_1,t_{10}]}$. To do this, it is necessary to compute all the minimal proofs of $\alpha^{[t_1,t_{10}]}$ in $\mathbb{K}$. In this case, $\Pi(\alpha^{[t_1,t_{10}]}, \mathbb{K}) = \{\{\beta^{[t_1,t_{10}]}, (\beta^{[t_1,t_5]} \rightarrow \alpha^{[t_1,t_5]})^{[t_1,\infty]}, (\beta^{[t_6,t_{10}]} \rightarrow \alpha^{[t_6,t_{10}]})^{[t_1,\infty]}\}\}$. Then, $S = \sigma^c(\alpha^{[t_1,t_{10}]}, \mathbb{K}) = \{(\beta^{[t_1,t_5]} \rightarrow \alpha^{[t_1,t_5]})^{[t_1,\infty]}, (\beta^{[t_6,t_{10}]} \rightarrow \alpha^{[t_6,t_{10}]})^{[t_1,\infty]}\}$. Thus, $\overset{out}{S}{}^{t_1}_{t_{10}} = \emptyset$. Therefore, $\mathbb{K} \otimes \neg\alpha^{[t_1,t_10]} = \{\beta^{[t_1,t_{10}]}, \delta^{[t_4]}, \neg\alpha^{[t_1,t_{10}]}\}$.*

5. Related work

Alchourrón and Makinson were the first to logically study the changes of a legal code [2,3,1]. The addition of a new norm *n* causes an enlargement of the code, consisting of the new norm plus all the regulations that can be derived from *n*. Alchourrón and Makinson distinguish two other types of change. When the new norm is incoherent with the existing ones, we have an *amendment* of the code: in order to coherently add the new regulation, we need to reject those norms that conflict with *n*. Finally, *derogation* is the elimination of a norm *n* together with whatever part of the legal code that implies *n*.

[4] inspired by the works above proposed the so called AGM framework for belief revision. This area proved to a very fertile one and the phenomenon of revision of logical theories has been thoroughly investigated. It is then natural to ask if belief revision offers a satisfactory framework for the problem of norm revision. Some of the AGM axioms seem to be rational requirements in a legal context, whereas they have been criticised when imposed on belief change operators. An example is the *success* postulate, requiring that a new input must always be accepted in the belief set. It is reasonable to impose such a requirement when we wish to enforce a new norm or obligation. However, it gives rise to irrational behaviors when imposed to a belief set, as observed in [11].

The AGM operation of contraction is perhaps the most controversial one, due to some postulates such as recovery [14,25], and to elusive nature of legal changes such as derogations and repeals, which are all meant to contract legal effects but in remarkably different ways [14]. Standard AGM framework is of little help here: it has the advantage

of being very abstract—it works with theories consisting of simple logical assertions—but precisely for this reason it is more suitable to capture the dynamics of obligations and permissions than the one of legal norms. In fact, it is hard in AGM to represent how the same set of legal effects can be contracted in many different ways, depending on how norms are changed. For this reason, previous works [12,13,14] proposed to combine a rule-based system with some forms of temporal reasoning.

Difficulties behind standard AGM have been considered and some research has been carried out to reframe AGM ideas within reasonably richer rule-based logical systems, combining AGM ideas with Defeasible Logic [23,15] or Input/Output Logic [7,24]. [25] suggested a different route, i.e., employing in the law existing techniques—such as iterated belief change, two-dimensional belief change, belief bases, and weakened contraction—that can obviate problems identified in [14] for standard AGM.

In this paper we showed to extend base revision with temporal reasoning, and, in particular, with time intervals. Our approach, like in [14], is able to deal with constituents holding in an interval of time, thus an expression $\Longrightarrow a^{[t_1,t_2]}$ meaning that a holds between t_1 and t_2 can be seen as a shorthand of the pair of rules from [14] (defeasible and defeater) $\Longrightarrow a^{[t_1,pers]}$ and $\rightsquigarrow \neg a$. We have taken this design decision because it simplifies the construction of the revision operator: following the semantics of the temporalised rule proposed in [14] and explained in Section 3 (an adapted version), the revision operator in many cases only consists in modifying the intervals to maintain the consistency.

Interval and duration based temporal defeasible logic have been developed [6,17]. [17] focuses on duration and periodicity and relationships with various forms of causality. [6] proposed a sophisticated interaction of defeasible reasoning and standard temporal reasoning (i.e., mutual relationships of intervals and constraints on the combination of intervals). In both cases it is not clear whether the techniques employed there are relevant to the application to norm modifications, and such works consider only a single temporal dimension.

In [8], belief revision in a temporal logic context is also addressed. However, they use modal operators over possible worlds to model belief changes. Here we are focused in a propositional language with time intervals following kernel contraction construction proposed in [19].

6. Conclusions and Future Work

In this work we have introduced a time-based belief revision operator for legal systems. A temporalised belief base and a temporalised belief derivation was defined, following the formalisation of temporal rules, suitable to model examples in the legal area. In this special belief base each piece of information is decorated with a time interval. In this scenario our novel belief revision operator allows the consistent addition of temporalised sentences in a temporalised belief base. If a contradiction arises, then the revision operator may either remove conflictive temporalised sentences or modify the intervals.

Change operators are presented following the AGM model [4] where the operators are defined through constructions and representation theorems. In this paper, the operator was defined through construction. As future work, rationality postulates will be given and its corresponding Representation Theorem for this new revision operator. This theorem proves the correspondence between the set of postulates and the construction.

Acknowledgments

This work was partially supported by PGI-UNS (grants 24/ZN30, 24/ZN32) and EU H2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 for the project MIREL: MIning and REasoning with Legal texts.

References

[1] Carlos E. Alchourrón and Eugenio Bulygin. The expressive conception of norms. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 95–125. D. Reidel, Dordrecht, 1981.
[2] Carlos E. Alchourrón and David C. Makinson. Hierarchies of regulations and their logic. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 125–148. D. Reidel, Dordrecht, 1981.
[3] Carlos E. Alchourrón and David C. Makinson. The logic of theory change: Contraction functions and their associated revision functions. *Theoria*, 48:14–37, 1982.
[4] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.
[5] James F. Allen. Towards a general theory of action and time. *Artif. Intell.*, 23(2):123–154, 1984.
[6] Juan Carlos Augusto and Guillermo Ricardo Simari. Temporal defeasible reasoning. *Knowl. Inf. Syst.*, 3(3):287–318, 2001.
[7] G. Boella, G. Pigozzi, and L. van der Torre. A normative framework for norm change. In *Proc. AAMAS 2009*, pages 169–176. ACM, 2009.
[8] Giacomo Bonanno. Axiomatic characterization of the AGM theory of belief revision in a temporal logic. *Artif. Intell.*, 171(2-3):144–160, 2007.
[9] Maximiliano Celmo Budán, Maria Laura Cobo, Diego C. Martínez, and Guillermo Ricardo Simari. Bipolarity in temporal argumentation frameworks. *Int. J. Approx. Reasoning*, 84:1–22, 2017.
[10] André Fuhrmann. Theory contraction through base contraction. *Journal of Philosophical Logic*, 20(2):175–203, may 1991.
[11] Dov M. Gabbay, Gabriella Pigozzi, and John Woods. Controlled revision - an algorithmic approach for belief revision. *J. Log. Comput.*, 13(1):3–22, 2003.
[12] G. Governatori, M. Palmirani, R. Riveret, A. Rotolo, and G. Sartor. Norm modifications in defeasible logic. In *JURIX 2005*, pages 13–22. IOS Press, Amsterdam, 2005.
[13] G. Governatori, A. Rotolo, R. Riveret, M. Palmirani, and G. Sartor. Variants of temporal defeasible logic for modelling norm modifications. In *Proc. ICAIL'07*, pages 155–159, 2007.
[14] Guido Governatori and Antonino Rotolo. Changing legal systems: legal abrogations and annulments in defeasible logic. *Logic Journal of the IGPL*, 18(1):157–194, 2010.
[15] Guido Governatori, Antonino Rotolo, Francesco Olivieri, and Simone Scannapieco. Legal contractions: a logical analysis. In *Proc. ICAIL 2013*, 2013.
[16] Guido Governatori, Antonino Rotolo, and Giovanni Sartor. Temporalised normative positions in defeasible logic. In Giovanni Sartor, editor, *Proc. ICAIL 2005*. ACM, 2005.
[17] Guido Governatori and Paolo Terenziani. Temporal extensions to defeasible logic. In *AI 2007*, pages 476–485, 2007.
[18] Sven Ove Hansson. In defense of base contraction. *Syntheses*, 91(3):239–245, june 1992.
[19] Sven Ove Hansson. Kernel Contraction. *The Journal of Symbolic Logic*, 59:845–859, 1994.
[20] Sven Ove Hansson. *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Kluwer Academic Publishers, 1999.
[21] H. L. A. Hart. *The Concept of Law*. Clarendon Press, Oxford, 1994.
[22] Hans Kelsen. *General theory of norms*. Clarendon, Oxford, 1991.
[23] Antonino Rotolo. Retroactive legal changes and revision theory in defeasible logic. In G. Governatori and G. Sartor, editors, *DEON 2010*, volume 6181 of *LNAI*, pages 116–131. Springer, 2010.
[24] Audun Stolpe. Norm-system revision: theory and application. *Artif. Intell. Law*, 18(3):247–283, 2010.
[25] Gregory R. Wheeler and Marco Alberti. No revision and no contraction. *Minds and Machines*, 21(3):411–430, 2011.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-59

Giving Every Case Its (Legal) Due

The Contribution of Citation Networks and Text Similarity Techniques to Legal Studies of European Union Law

Yannis PANAGIS [a,1], Urška ŠADL [a,b] and Fabien TARISSAN [c]
[a] *iCourts, Centre of Excellence for International Courts*
[b] *European University Institute*
[c] *Université Paris-Saclay, ISP, ENS Paris-Saclay, CNRS*

Abstract. In this article we propose a novel methodology, which uses text similarity techniques to infer precise citations from the judgments of the Court of Justice of the European Union (CJEU), including their content. We construct a complete network of citations to judgments on the level of singular text units or paragraphs. By contrast to previous literature, which takes into account only explicit citations of entire judgments, we also infer implicit citations, meaning the repetitions of legal arguments stemming from past judgments without explicit reference. On this basis we can differentiate between different categories and modes of citations. The latter is crucial for assessing the actual legal importance of judgments in the citation network. Our study is an important methodological step forward in integrating citation network analysis into legal studies, which significantly enhances our understanding of European Union law and the decision making of the CJEU.

Keywords. Network analysis, Citation networks, Text similarity, CJEU

1. Introduction

While citation network analysis has gained traction as an approach to understand law and courts, legal scholars remain reserved. Our study is motivated by this reticence, which can be summed up in three pertinent objections. We discuss them in turn. First, legal scholars see the approach as quantitative hence unfit for detailed qualitative investigations of legal rules and principles and their application to concrete cases. The latter are typically considered as the main purpose of a legal study. The position is reinforced by the existing use of the approach since, bar a few exceptions, the network approach has been paired with statistical quantitative analysis. The field is strongly focused on judicial behavior and judicial bias in legal decision making, using case citation networks to answer questions related to law rather than legal questions. This is particularly recognizable in research on the United States Supreme Court (USSC) but has been less true for studies of the CJEU [7]. Examples include inquiries into judicial activism, the rise of *stare decisis*, the depreciation of precedents in the USSC [4], citation strategies of international courts, and the (strategic) behavior of individual judges [8].

[1]Corresponding Author, email: `ioannis.panagis@jur.ku.dk`. This research is partly funded by the Danish National Research Foundation Grant no. DNRF105.

Second, the network analysis approach treats all citations as equally important, and does not discriminate between different types of references. The judge might mention a case in passing or include it in a string citation; she might cite it to distinguish it from the case at hand or dismiss it as irrelevant, or because one of the parties to the case relied on it. She might, furthermore, cite the case as an example, to reason by analogy, rather than employ it as a binding, guiding, or even legally persuasive source of law that legally or *de facto* obliges her to reach a specific outcome. The information is crucial for the inquiry of what is valid law and which cases are the truly important reference points in a court's repository. The criticism is underscored by the fact that so far existing studies using case citations have in fact conceptualized citations as equal, and treated them as legally relevant. Moreover, the same studies have assumed that the "citation behavior of the Court provides information about which precedents serve important roles in the development of [. . .] law," [6]. The implication of this assumption would be that "[. . .] a judgment's value as a source of law is limited if it has never been cited by the Court of Justice." [7], which is something that most legal scholars would contest.

Third, because citation network analysis relies on explicit case citations *it can only be applied to courts with a developed and rigorous citation practice*. The criticism raises the question whether the method will yield inaccurate findings in the case of continental style courts like the CJEU, which tend to repeat the wording and the arguments established in past cases without citing the source (the so-called *implicit citations*). Albeit this is less true for the more recent judgments, the CJEU has especially in its earlier cases often resorted to such implicit citations.

To address the above challenges we combine the network approach with text analysis. First, we construct the citation network based on references to paragraphs of individual judgments as units rather than judgments as a whole (we use cases and judgments of the CJEU as synonyms). Namely, most cited paragraphs typically include a particular concept or a particular formulation, which is relevant in the process of construction of legal arguments[2]. By doing this we take a step further in identifying the aspects of cases that are legally important. We also acquire the information whether the case is relevant for one or several legal aspects. For instance, if only one paragraph of a case is repeatedly cited, the case is most likely important for resolving one legal issue. If, by contrast, there are several different paragraphs of one judgment that are cited, the case might be important for resolving more than one legal issue.

Second, we isolate the references that are directed to entire cases (global references) and use text similarity techniques to infer local references, references to particular parts of cases (paragraphs). The latter are called *implicit references* or *missing citations*, where the CJEU repeats the text of a particular paragraph / part of the judgment verbatim or with slight variations but does not cite it. Third, we assess the relevance of cited paragraphs from a legal perspective. We showcase our approach by evaluating the links to three of the best known cases in the CJEU doctrine: *Dassonville*, *Defrenne II* and *Francovich*[3].

[2]Typically, the judgments of the CJEU are separated into self-contained units or paragraphs, dealing with a particular point of law or fact. In the older judgments dating back to the 1970s the paragraphs are not numbered systematically. Later, in the 1980s, when the judgments became longer and the writing style of the CJEU more argumentative and informative, the CJEU began to number the paragraphs.

[3]Dassonville: case C-8/74, ECLI:EU:C:1974:82, Defrenne II: case C-43/75, ECLI:EU:C:1976:39 (also known as "Defrenne II") and Francovich: Case C-6/90, ECLI:EU:C:1991:428.

Table 1. The different types of references, illustrated on references to *Dassonville*.

Reference type	Paragraph
local	39. The prohibition of measures [...] (see, in particular, **Case 8/74 Dassonville [1974] ECR 837, *paragraph 5***; Case 178/84 Commission v Germany [1987] ECR 1227 ("Beer purity law"), paragraph 27; [...].
global	10 It must be recalled first of all that since its judgment in **Case 8/74 Procureur du Roi v Dassonville [1974] ECR 837**, the Court has consistently held [...]

To summarize, by leveraging the fine-grained paragraph data we: a) infer the missing citations and b) tease out and assess the potentially legally relevant parts of the cited case. Altogether, this information is crucial to evaluate the actual legal importance of a particular case and its influence on the development of legal doctrine.

The rest of the paper is organized as follows: in Section 2 we present the methods that we use, namely the missing link detection technique, lay out the assumptions and observations, on which we base our research strategy and explain the terminology (judicial formulas). In Section 3 we present a quantitative and qualitative evaluation of our the missing link detection technique and, finally, we conclude in Section 4.

2. Research Strategy, Method and Data

Our research strategy is based on a set of assumptions and observations about the CJEU and its style of decision writing.

2.1. Paragraphs, Networks, and Reference Types

Every judgment of the CJEU is divided into smaller text units or paragraphs. The paragraphs form the skeleton of the judgment and contain the legal arguments that the Court is communicating as well as references to previous cases (precedents), on which the CJEU relies in order to support these arguments.

We define a *judgment paragraph* as the part of a judgment that usually starts with an integer number, e.g. 10, and extends until the text paragraph starting with the next integer, i.e. 11, excluding perhaps quoted text. References (or citations) can be grouped into two categories: *local* references that precisely define which paragraph of the previous judgment is being cited with a number and *global* references where the entire judgment is cited without specifying the paragraph(s). Examples of both types are given in Table 1.

A citation network is defined as a pair $G = (V, E)$, where $V = V_{case} \cup V_{par}$ is the set of nodes, V_{case} is a set of cases that is referred to by means of global references, V_{par} is the set of judgment paragraphs that cite and get cited by means of local references, $E = E_{global} \cup E_{local}$ are the edges such that $E_{global} = \{(u,v) | u \in V_{par}, v \in V_{case}\}$ depict the global citations and $E_{local} = \{(u,v) | u,v \in V_{par}\}$ are the local ones. Lastly, we denote by $par(C)$ the set of all judgment paragraphs of a given case C.

2.2. Formulas

All language users depend on prefabricated phrases. That said, the language of courts is formalized to a much larger extent than natural language and is by far more repetitive. The language of the CJEU is particularly routinized, even when compared to the CJEU's

counterparts in France, Germany and the United Kindgom [1,14]. The CJEU makes use of a limited set of textual devices to construct its arguments [14]. These have been labeled judicial *formulas* in literature [1] and can be defined as legal phrases, which the CJEU repeats as self-standing statements of the law or in context with other prefabricated phrases. The formulas are not only rhetorical but simultaneously characterize the European Union legal order, establish its principles and fundamental concepts [1]. They speed up the process of judgment writing and make searching for relevant (legally similar) past cases more effective.[4]

With repetition the formulas detach from the judgments in which they were first pronounced (the original judgments) and acquire a broader relevance. They begin to function as abstract rules [11]. The modification of the content of the formulas reflects how the CJEU develops, elaborates, expands, or restricts legal concepts, and the reach of European Union law and how it adapts broad formulations to fit individual situations [11, 2]. This does not imply that the original judgment loses its legal relevance because it is not cited but rather that the legal relevance of the judgment becomes embedded, or implicitly acknowledged [10].

Among the best known examples is the so-called *Van Gend* formula, where the CJEU defined the Treaty as establishing *"a new legal order of international law"* and the formula in the *Grzelczyk* judgment[5], where the CJEU defined the concept of European Union citizenship, stating that *"Union citizenship is destined to be the fundamental status of nationals of the Member States."*. Both had far-reaching implications for the relationship between the European Union and the Member States and between the European Union and other international organizations.

2.3. Text Processing and Text Similarity

As already pointed out in the previous section the CJEU often paraphrases the wording of the original formula to express the same content. The new versions of the formula are thus not identical but similar to the original formula. To infer implicit citations (missing links) and local references we thus rely on text similarity.

We proceed by first applying a typical Natural Language Processing workflow which consists of the following steps: a) sentence and word segmentation, b) lowercasing, c) stemming, d) removal of stopwords, single letter words and numbers. For the last step we use the standard list of English stopwords.

Our purpose is to define a way to measure the formula similarity between any mutated paragraph a and the original paragraph b, and use this metric to detect the actual cited paragraphs in the case of global references. We therefore, use a special case of the Tversky index [15] (see Equation 1).

$$T(b,a) = \frac{|b \cap a|}{|b \cap a| + \alpha|b-a| + \beta|a-b|} \qquad \alpha,\beta \geq 0 \tag{1}$$

where $|\cdot|$ here denotes the number of words.

The *formula similarity index* between paragraphs b and a, $fsi(b,a)$, is merely the value of Tversky index we get by substituting $\alpha = 1$ and $\beta = 0$ and thus eliminating the influence of the mutant paragraph to the similarity score, which is desirable. Hence:

[4]Scholars have called these pre-fabricated phrases the *building blocks*, see e.g. [3]

[5]The corresponding ECLI numbers are ECLI:EU:C:1962:42 (Van Gend) ECLI:EU:C:2001:458 (Grzelzyk)

$$fsi(b,a) = \frac{|b \cap a|}{|b|} \tag{2}$$

Note that the above definition is not symmetric, i.e. $fsi(a,b) \neq fsi(b,a)$. An interesting property of the proposed similarity index is that it implies that the paragraphs are treated as *bags-of-words*, in the sense that the order of the words is not important. The latter property, together with stemming help us to partly overcome the effect of paraphrasing during link detection. In the context of inferring implicit links, a given paragraph a, refers to case B but not to a specific paragraph $b \in par(B)$. Hence, we will use *fsi* to infer which paragraph of B should have been the target of the implicit reference.

2.4. Dataset Construction

The dataset for this paper was first compiled by downloading from EUR-Lex[6], the texts of all judgments of the CJEU until the end of 2015. This yielded 10418 documents (judgments) in total. We then extracted all paragraphs in the *Grounds* section of the judgments[7]. We kept the English language versions of the judgments whenever available, and supplemented the dataset with the texts of the French language versions, yielding a total of around 445000 paragraphs. Since the CJEU did not number judgment paragraphs systematically until the 1970s, we excluded all judgments that do not have numbered paragraphs from the extraction process.

Note that due to this technical issue, some older cases are left out in the present study. This includes some important landmark cases such as *Van Gend* and *Costa* for instance. However, we argue that this does not affect the way we assess the relevance of our methodology in Section 3, which is the core of the present study. This rather raises the question of completing the dataset which we let for a future work.

Subsequently, we employed the core extraction methodology in [10], i.e. used GATE and a set of JAPE rules [5], to infer citations to paragraphs and to build a paragraph-to-paragraph citation network out of the entire paragraph dataset, including both *global* and *local* references. The main difference from the core methodology proposed in [10] is that in order to annotate the case names in the text, where possible, we preprocessed the paragraphs before passing them to GATE, instead of using a gazetteer.

Preprocessing was a necessary step to identify citations in the text that refer to the case by the name that it is commonly known by, in CJEU. For instance, the CJEU very often refers to a judgment without using case numbers, like in the following: *"37. The Court stated in* ***paragraph 16 of Keck and Mithouard****, cited above, that national provisions [...] within the meaning of the line of case-law initiated by* ***Dassonville****, cited above"*[8]. Text fragments like the previous, can however be annotated with the CELEX number of the case, by using a white-list of case names. The annotation can then be further used to identify every single case decided by the CJEU and stored in EUR-Lex. The use of gazetteer reaches the same final result in the general case, makes things more complicated, however, in the presence of ambiguous citations, e.g. *"Commision v. France"*.

The key figures of both the paragraph-to-paragraph and the corresponding case-to-case networks are summarized in Table 2.

[6]http://eur-lex.europa.eu

[7]The judgments of the CJEU are divided in sections. The section Grounds contains the statements of the CJEU about the legal arguments and is thus the part of the judgments that is most relevant for legal scholars.

[8]Karner, Case C-71/02, ECLI:EU:C:2004:181, par. 37

Table 2. The case-to-case and the paragraph-to-paragraph network of the CJEU. The numbers in parenthesis indicate, $|V_{case}|$ for nodes and $|E_{global}|$ for edges, respectively.

	case-to-case	paragraph-to-paragraph
Nodes	10418	74219 (4773)
Edges	49519	93713 (18778)

2.5. Predicting Local Links

The paragraph-to-paragraph network and the global references open the possibility to complement the network by predicting the actual target paragraphs from global citations on the basis of citing paragraphs alone. While legal experts who are familiar with the judgments of the CJEU and the relevant formulas would see this as an intuitive step in the legal analysis, the task is less straightforward for a computer program.

We nonetheless designed the following simple algorithm to overcome this difficulty: For every edge $(p,C) \in E_{global}$, run through the set $par(C)$ and for every paragraph $p_c \in par(C)$, compute $fsi(p_c,p)$ from Equation 2. We compute a candidate target paragraph p_t, taking $p_t = \max\{fsi(p_c,p) \forall p_c \in par(C)\}$ and then check if $fsi(p_t,p) \geq t$, for a specified threshold t, in which case we add the edge (p,p_t), which means that we infer that p should cite p_t among all paragraphs of C. If $fsi(p_t,p) < t$, no paragraph is predicted.

Computing the score $fsi(p_c,p)$, as above, implies that we consider p_c as the (candidate) source of law that paragraph p is citing, and *fsi* represents the percentage of p_c repeated by p. Another implication of the above approach is that we predict at most one edge for every (p,C)-pair even though in principle, a paragraph could refer to more than one paragraph of a cited judgment.

The selection of an appropriate value for t is not straightforward. In our case we worked with different values of t and observed that selecting $t \geq 0.5$ would exclude several true positive citations, when the formula that was reproduced in the global citation was a rather small fraction of the original formula. In fact we calculated the average $fsi_{avg} = \{fsi(v,u)/|E_{local}|, \forall (u,v) \in |E_{local}|\}$ and the result was $fsi_{avg} = 0.48$. Therefore, we tested several values of $t < 0.5$ and we ended with $t = 0.4$, which we will use for the rest of the paper. We omit the full results for a longer version of the paper.

3. Results and Interpretation

3.1. Predicting Local Citations

In order to assess the prediction method we evaluated the quality of the predictions by examining the predicted links towards three *landmark cases*[9] of the case-law of the CJEU, *Dassonville, Defrenne II* and *Francovich*. The reason for selecting those cases is on the one hand their qualitative characteristics, in particular their perceived doctrinal difference and versatility, and then their quantitative characteristics, see Table 3. As we see from Table 3, the three selected cases are cited more on average and vary greatly in the number of paragraphs of them that get cited.

[9] see e.g. [13] for the discussion on landmark cases

Table 3. Summary statistics for the number of citations of the selected cases compared with the network all paragraphs. The number of cited paragraphs for the entire network is the average.

	Dassonville	Defrenne II	Francovich	Network
Median	1.00	3.00	2.00	1.00
Mean	33.00	3.53	4.00	2.12
Number of cited paragraphs	3	19	27	5.23

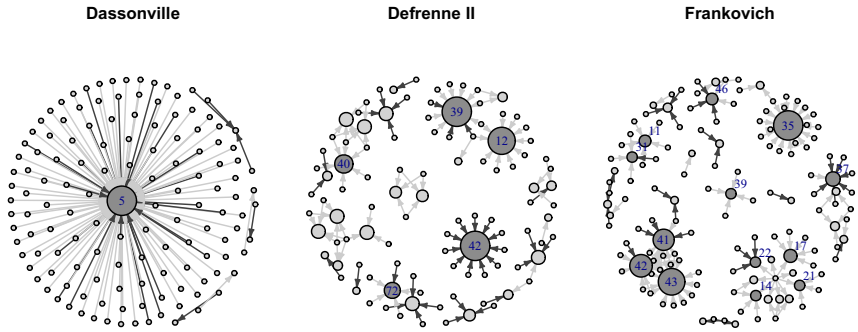


Figure 1. Examples or paragraph networks. Dark edges denote predicted links and numbered nodes correspond to highly cited paragraphs of each case.

Dassonville is prominent among legal scholars with regard to one legal aspect in a well defined area of EU law (free movement of goods). Defrenne II is typically considered by legal scholars in several areas of EU law, i.a. for its contribution to the general principle of non-discrimination on grounds of gender (the principle of equal pay for equal work), the limitation of temporal effects of judgments of the CJEU in exceptional circumstances, and horizontal direct effect of the Treaty as a fundamental characteristics of the EU legal order. Dassonville and Defrenne II are cases of creative judicial interpretation of the Treaty as the principal, written legal source of EU law. By contrast, Francovich is known as a judicial innovation. It establishes a new legal principle that does not originate from a written legal source of EU law and lays down the conditions under which it can be applied. This is reflected in the high number of different paragraphs that are cited in subsequent cases. The induced subgraphs of the paragraph networks of the above three cases are juxtaposed in Figure 1, where their differences in legal substance are very nicely represented by the fact that the Dassonville subnetwork consists almost entirely of one star, with incoming citations only to par. 5, whereas in Defrenne II and Francovich the citations to several legal aspects produce a number of smaller clusters.

This implies first, that Dassonville contains one formula, which is most often repeated in subsequent cases as a whole. By contrast, Defrenne II and Francovich contain more than one formulas. Second, since the Dassonville formula has a single and distinct meaning related to a particular legal problem, it is consistently repeated in one particular legal and several, factually similar contexts. Defrenne II and Francovich have a broader legal relevance because they concern the basic principles of the EU legal order that are applicable across subject areas. They can thus be repeated in more than one legal context and to factually distinct situations.

Altogether, we predicted 97 citations to all three cases while 33 citations remained unmatched. A legal expert validated the approach by reviewing the list of predicted citations and determining whether the citations were accurate. As *accurate citations*, we

Table 4. Method evaluation

Case	Recall	Precision	F1	Baseline Precision
Dassonville	87.1%	90.0%	88.5%	85.2%
Defrenne II	49.0%	66.7%	56.5%	35.6%
Francovich	85.7%	77.4%	81.4%	0.0%
TOTAL	69.4%	77.3%	73.2%	40.5%

considered citations that either matched the cited paragraph on legal language level (*text match*), meaning that they repeated the words of the formula, and on legal content level, meaning that they were predicted in a meaningful legal context and hence legally relevant (*content match*), or both (*full match*). Table 4 presents the results of this evaluation against a baseline approach, and with regard to *Precision*, *Recall* and *F1* measures, see [9]. Table 4 shows the performance both per case and in total (last row).

The baseline for our comparisons was to assign the local link to the most cited paragraph of the case. This method yielded a total precision of 40.5%. By way of comparison, the formula similarity index method was almost two times better with 77.3% precision with a satisfactory recall of 69.4% for all three cases.

3.2. Various Types of Citations

The least satisfactory results concern cases, in which the formulas are – in part or whole – repeated several times within a single document, as in Francovich, pars. 28, 37 and 46. For instance, a formulation that the Member States are "*obliged to make good loss and damage suffered by individuals as a result of the failure to transpose...*", appears in all three paragraphs, however, for very different reasons: in par. 28 as a reproduction of the question of the national court ("*28. In the second part of the first question the national court seeks to determine whether a Member State is obliged to make good loss and damage suffered by individuals as a result of the failure to transpose Directive 80/987*"), in par. 37 as a genuine statement of the law by the CJEU ("*37. It follows from all the foregoing that it is a principle of Community law that the Member States are obliged to make good loss and damage caused to individuals by breaches of Community law for which they can be held responsible.*") and in par. 46 as the reply of the CJEU to the national court ("*46 The answer to be given to the national court must therefore be that a Member State is required to make good loss and damage caused to individuals by failure to transpose Directive 80/987.*"). While the reference to par. 46 is recalled with a 100% precision, the reference to par. 28 is recalled with 0% precision and the reference to par. 37 is recalled with 87% precision[10]. In cases of high inter-textual similarity such as in our example the first occurrence would most likely be a reformulation of a question, while the last occurrence would most likely be an answer. The central – both legally and with regard to the position in the text – would be the middle occurrence.

Generally speaking, our predictions failed mostly with regard to linguistically too indistinct formulations and longer formulations, which repeated the arguments of the parties, or the questions of the national courts that also repeated the formulas taken from past cases, often tying them to the particular facts of the case, or in combination with either national or EU secondary legislation. The confusion arose because the CJEU refers to these arguments or preliminary questions in the Grounds of the judgment, to indicate

[10]Due to space constraints we have omitted the detailed results from this paper.

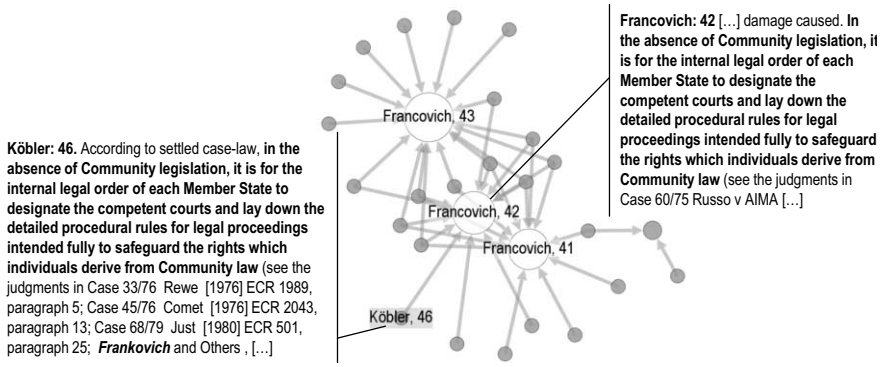


Figure 2. A predicted link from Köbler, 46 to Francovich, 42. The repeated formula is shown with bold letters.

which one of the several legal issues it is dealing with. The latter is especially common in longer and legally more complex judgments, where the national court formulates several preliminary questions.

3.3. Revealing How a Case Is Used

The most interesting findings concern individual predictions. For instance, a legally novel development of the so-called *Francovich principle* occurred first in *Brasserie du Pecheur*, and later in *Köbler*. While the workflow outlined in Section 2.4 does not detect a citation from *Köbler*, par. 46, to *Francovich* par. 42, due to a typo, a misspelling of "Francovich", as shown in Figure 2, the citation is detected on the basis of text similarity. This is not an isolated occurrence hence it can be argued that the text similarity techniques are indispensable for obtaining a more accurate picture of case citations and case centrality.

The final example demonstrates the contribution of our approach to the study of legal development, in particular legal change. Namely, the approach, which the CJEU created with *Dassonville* with regard to the national measures restricting trade, was importantly narrowed down in *Keck*. The central paragraph, in which this occurs, is *Keck*, par. 16, which refers to *Dassonville*, par. 5. Our method successfully detects this reference by link prediction on the basis of text similarity. The analysis on document level, without the use of text similarity would not detect this reference, which is crucial for legal scholars.

4. Conclusions

In this article we constructed a network of individual text units or paragraphs of the judgments of the CJEU and used text similarity techniques to obtain a complete information about the content of case citations. This level of granularity enabled us to draw a more complete picture of implicit citations, meaning the repetitions of legal arguments stemming from past judgments without explicit references to those judgments. The implicit citations provided the missing data about the actual use of a case by the CJEU. On the basis of the precise information about the content of citations we were furthermore able to differentiate between various types of citations. Together, this information is important to empirically determine to what extent a specific case has influenced the law, and thus giving every case its doctrinal due.

Our findings show first, that algorithms can correctly predict the local links (citations to specific paragraphs), in cases where the formulations are limited to a one-sentence linguistically characteristic original sequence. By contrast, the results were not as convincing in the case of very short or very long linguistically indistinct formulations with broad legal application (for instance, a reference to "*legal certainty*").

Second, the findings reveal that it is possible to tease out legal development by a more detailed categorization of predicted links. A content match would often indicate a mutation of formulas, which is often an indication of legal change or important legal innovation (this was the example of Keck, par. 16, citing Dassonville, par. 5, only to reverse the course of the law established by Dassonville).

Our study is, to the best of our knowledge, the first to demonstrate that by combining the citation network approach with text similarity detection techniques, we can access the legal content behind citations. Thereby, our research opens avenues for original research, which by further improving and fine tuning the basic approach, can detect the doctrinal origin of legal formulas, their modifications in the judgments of the CJEU over time, as well as, their generalization across different areas of law. Finally, we believe that the findings of this paper will allow us to develop machine learning approaches to the problem of detecting legal formulas, in a spirit similar to recent developments, e.g. [12].

References

[1] Loïc Azoulai. La fabrication de la jurisprudence communautaire. *Dans la Fabrique du Droit Européen, Brussels: Bruylant*, pages 153–170, 2009.

[2] Loïc Azoulai. The retained powers' formula in the case law of the european court of justice: EU Law as Total Law. *Eur. J. Legal Stud.*, 4:178, 2011.

[3] Gunnar Beck. *The Legal Reasoning of the Court of Justice of the EU*. Bloomsbury Publishing, 2013.

[4] Ryan C Black and James F Spriggs. The citation and depreciation of US Supreme Court precedent. *Journal of Empirical Legal Studies*, 10(2):325–358, 2013.

[5] Hamish Cunningham. GATE, a general architecture for text engineering. *Comput Humanities*, 36(2):223–254, 2002.

[6] James H Fowler and Sangick Jeon. The authority of Supreme Court precedent. *Soc. networks*, 30(1):16–30, 2008.

[7] Johan Lindholm and Mattias Derlén. The court of justice and the Ankara agreement: Exploring the empirical approach. *Europarättslig tidskrift*, (3):462–481, 2012.

[8] Yonatan Lupu and James H Fowler. Strategic citations to precedent on the us supreme court. *The Journal of Legal Studies*, 42(1):151–186, 2013.

[9] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[10] Yannis Panagis and Urška Šadl. The force of EU case law: A multi-dimensional study of case citations. In *JURIX*, pages 71–80, 2015.

[11] Urška Šadl. Case–Case-Law–Law: Ruiz Zambrano as an illustration of how the court of justice of the European Union constructs its legal arguments. *Eur Const Law Rev*, 9(2):205–229, 2013.

[12] Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, 25(1):107–126, Mar 2017.

[13] Fabien Tarissan, Yannis Panagis, and Urška Šadl. Selecting the cases that defined Europe: complementary metrics for a network analysis. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016)*, pages 661–668. IEEE, 2016.

[14] Aleksandar Trklja. A corpus investigation of formulaicity and hybridity in legal language: a case of EU case law texts: A case of EU case law texts. In S.G. Roszkowski and G. Pontrandolfo, editors, *Phraseology in Legal and Institutional Settings: A Corpus-based Interdisciplinary Perspective*. Routledge, 2017.

[15] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-69

Argument Schemes for Discussing Bayesian Modellings of Complex Criminal Cases

Henry PRAKKEN [1]
Utrecht University and University of Groningen

Abstract. In this paper two discussions between experts about Bayesian modellings of complex criminal cases are analysed on their argumentation structure. The usefulness of several recognised argument schemes is confirmed, two new schemes for interpretation arguments and for arguments from statistics are proposed, and an analysis is given of debates about the validity of arguments. From a practical point of view the case study yields insights into the design of support software for discussions about Bayesian modellings of complex criminal cases.

Keywords. Argument schemes, reasoning about evidence, Bayesian probability theory, argumentation support

1. Introduction

There is an ongoing debate on what is the best model of rational evidential reasoning in criminal cases. Both argumentation-based, story-based and Bayesian approaches have been proposed [3]. In this paper I remain neutral with respect to this debate. Instead I will argue that even if a Bayesian approach is adopted, there is still one clearly argumentative aspect of this form of reasoning, namely, debates about the merits of a proposed Bayesian model. This observation is theoretically interesting but also has practical implications for support systems for legal proof and crime investigation. Forensic experts increasingly use Bayesian probability theory as their theoretical framework and they increasingly use software tools for designing Bayesian networks. In crime investigation or in court the need may arise to record the pros and cons of the various design decisions embodied in the experts' analyses, and argumentation support technology may be of use here.

To obtain insight in the requirements for argumentation-based add-ons to Bayesian-network software tools, this paper examines two recent Dutch criminal cases in which I was appointed by courts to comment on a Bayesian analysis of the entire case proposed by an expert of the prosecution. In the present paper I analyse to what extent our expert reports and written replies contain arguments that can be classified as instances of argument schemes or as applications of critical questions of these schemes.

[1] Department of Information and Computing Sciences, Utrecht University, and Faculty of Law, University of Groningen, The Netherlands; E-mail: H.Prakken@uu.nl.

2. The cases

In the *Breda Six* case three young men and three young women were accused of killing a woman in the restaurant of her son after closing time, in 1993. The six were initially convicted in two instances, mainly on the basis of confessions of the three female suspects. In 2012 the Dutch Supreme Court reopened the case because of doubts about the truthfulness of these confessions. After a new police investigation the six were tried again by the court of appeal of The Hague and in 2015 they were again all found guilty, mainly on the ground that new evidence had confirmed the reliability of the confessions.

The prosecution in the case brought in an 80 page expert report by the climate physicist Dr. Alkemade (henceforth 'A') containing a Bayesian analysis of the entire case. A claimed that he could give a Bayesian analysis of the case since he had experience with using Bayesian probability theory in his work as a climate physicist. In his report, he concluded that on the basis of the evidence considered by him the probability that at least one of the six suspects was involved in the crime was at least 99,7%. The investigating judge in the case asked me to assess and evaluate A's report, which I did in a 41 page report. My main conclusion was that A's claims had no objective basis. In its final verdict, the court ruled that A could be regarded as an expert for the purpose of the case but that his method cannot be regarded as a reliable method for analysing complex criminal cases, for which reason A's conclusions had to be disregarded.

In the *Oosterland* case a person was accused of being responsible for 16 small arson cases in the small town of Oosterland in a six-month period in 2013. Initially the suspect was acquitted, mainly on the grounds that the two main witness testimonies were unreliable. In the appeal case the prosecution again brought in a report by A, this time 79 pages long. A concluded that on the basis of the evidence considered by him the probability that the suspect was involved in at least a substantial number of the arson cases was at least 99,8%. The investigating judge in the appeal case asked me to asses the reliability of A's method and its application to the case. I delivered a 42 page report with essentially the same conclusions as in the Breda Six case. A then wrote a 47 page reply to my report, after which I wrote a 9 page reply to his reply. In 2016 the court of appeal convicted the suspect of 7 arson cases and acquitted him of the remaining 9 cases. The court stated that it had chosen to disregard A's report "considering" my criticism.

3. Theoretical background

Probability theory [2] defines how probabilities between 0 and 1 (or equivalently between 0% and 100%) can be assigned to the truth of statements. As for notation, $Pr(A)$ stands for the unconditional probability of A while $Pr(A \mid B)$ stands for the conditional probability of A given B. In criminal cases we are interested in the conditional probability $Pr(H \mid E)$ of a hypothesis of interest (for instance, that the suspect is guilty of the charge) given evidence E (where E may be a conjunction of individual pieces of evidence). For any statement A, the probabilities of A and $\neg A$ add up to 1. The same holds for $Pr(A \mid C)$ and $Pr(\neg A \mid C)$ for any C. Two pieces of evidence E_1 and E_2 are said to be statistically independent given a hypothesis H if learning that E_2 is true does not change $Pr(E_1 \mid H)$, i.e., if $Pr(E_1 \mid H \wedge E_2) = Pr(E_1 \mid H)$. The axioms of probability imply that such independence is symmetric. The axioms also imply the following theorems (here given in odds form). Let $E_1, \ldots, E_n$ be pieces of evidence and H a hypothesis. Then:

$$\frac{Pr(H \mid E_1 \wedge \ldots \wedge E_n)}{Pr(\neg H \mid E_1 \wedge \ldots \wedge E_n)} = \frac{Pr(E_n \mid H \wedge E_1 \wedge \ldots \wedge E_{n-1})}{Pr(E_n \mid \neg H \wedge E_1 \wedge \ldots \wedge E_{n-1})} \times \ldots$$
$$\ldots \times \frac{Pr(E_2 \mid H \wedge E_1)}{Pr(E_2 \mid \neg H \wedge E_1)} \times \frac{Pr(E_1 \mid H)}{Pr(E_1 \mid \neg H)} \times \frac{Pr(H)}{Pr(\neg H)}$$

This formula is often called the *chain rule* (in odds form). The fractions on the extreme right and left are, respectively, the *prior* and *posterior odds* of H and $\neg H$. Given that probabilities of H and $\neg H$ add up to 1, the *prior*, respectively, *posterior probability* of H can be easily computed from them. If all of $E_1, \ldots, E_n$ are statistically independent from each other given H, then the chain rule reduces to

$$\frac{Pr(H \mid E_1 \wedge \ldots \wedge E_n)}{Pr(\neg H \mid E_1 \wedge \ldots \wedge E_n)} = \frac{Pr(E_n \mid H)}{Pr(E_n \mid \neg H)} \times \ldots \times \frac{Pr(E_1 \mid H)}{Pr(E_1 \mid \neg H)} \times \frac{Pr(H)}{Pr(\neg H)}$$

which is *Bayes' theorem* (in odds form). This is the formula used by A in his reports. Its attractiveness is that to determine the posterior odds of a hypothesis, it suffices to, respectively, multiply its prior odds with the so-called likelihood ratio, or evidential force, of each piece of evidence. For each piece of evidence E_i all that needs to be estimated is how much more or less likely E_i is given H than given $\neg H$. If this value exceeds (is less than) 1, then E_1 makes H more (less) probable.

Elegant as this way of thinking is, it is usually not applicable since often the global independence assumption concerning the evidence is not justified. Hence the name *naive Bayes*. The more general chain rule is often also practically infeasible, because of the many combinations of pieces of evidence that have to be considered. As a solution, *Bayesian networks* have been proposed, which graphically display possible independencies with directed links between nodes representing probabilistic variables. For each value of each node, all that needs to be estimated is its conditional probability given all combinations of all values of all its parents. Evidence can be entered in the network by setting the probability of the value of the corresponding node to 1, after which the probabilities of the values of the remaining nodes can be updated.

Argumentation is the process of evaluating claims by providing and critically examining grounds for or against the claim. *Argument schemes* [6] capture typical forms of arguments as a scheme with a set of premises and a conclusion, plus a set of critical questions that have to be answered before the scheme can be used to derive conclusions. If a scheme is deductively valid, that is, if its premises guarantee the conclusion, then all critical questions of a scheme ask whether a premise is true. If a scheme is defeasibly valid, that is, if its premises create a presumption in favour of its conclusion, then the scheme also has critical questions pointing at exceptional circumstances under which this presumption is not warranted. In formal approaches to argumentation, such as *ASPIC*$^+$ [5], argument schemes are often formalised as (deductive or defeasible) inference rules and critical questions as pointers to counterarguments. In the present paper argument schemes and their critical questions will be semiformally displayed, where critical questions asking whether the premises of the scheme are true will be left implicit.

4. The case study

In this section I discuss arguments from the written expert reports, the written replies and (when relevant) the verdicts that can be classified as instances of argument schemes or as

applications of critical questions of these schemes. Most of the schemes are taken from the literature but in two cases a new scheme will be proposed.

4.1. Text interpretation arguments

Some arguments are interpretation arguments, since they interpret the natural-language text of an expert report. In [6] two schemes for arguments from vagueness, respectively, arbitrariness of verbal classification are given, meant for criticising vagueness or arbitrariness in an argument. In the present case studies no such criticism was expressed but nevertheless issues arose concerning the correct interpretation of fragments of the reports. This gives rise to a new scheme of **Arguments from text interpretation**:

E says "P"
P means Q

E asserts that Q

This argument seems deductively valid (indicated by the single horizontal line) so it can only be criticised on its premises. Usually only the second premise will be controversial. In my reports I used this scheme several times as an introduction to an argument against Q. In one case, A convinced me in a private conversation afterwards that he had meant something else, after which I retracted my argument against Q.

4.2. Arguments from expert opinion

An obviously relevant scheme for modelling expert testimony is **arguments from expert opinion**. This especially holds for Bayesian modellings, since expert judgement is a recognised source of subjective probabilities. The following version of the scheme is modelled after [6].

E is an expert in domain D
E asserts that P
P is within D
=====
P

The double horizontal line indicates that the scheme is presumptive. Therefore, the scheme has **critical questions** concerning exceptions to the scheme: (1) How credible is E as an expert source? (2) Is E personally reliable as a source? (3) Is P consistent with what other experts assert? (4) Is E's assertion of P based on evidence? Question (1) is about the level of expertise while question (2) is about personal bias.

In probability theory sometimes a sharp distinction is made between frequentist (objective) and epistemic (subjective) Bayesian probability theory. Probabilities based on frequencies as reported by statistics would be objectively justified, while probabilities reflecting a person's degrees of belief would be just subjective. However, his sharp distinction breaks down from both sides. To start with, selecting, interpreting and applying statistics involves judgement, which could be subjective. Moreover, a person's degrees of belief could be more than just subjective if they are about a subject matter in which s/he is an expert. The same holds for the judgements involved in applying frequency information and statistics: if made by someone who is an expert in the problem at hand, these judgements may again be more than purely subjective. So the issue of expertise is crucial in both 'objective' (frequentist) and 'subjective' (epistemic) Bayes.

In the two cases, the question whether the scheme's first premise is true was very relevant. In this respect the cases highlight the importance of a distinction: *P* can be a specific statement made by the expert about a specific piece of evidence but it can also be a collection of similar statements or even the entire expert report. What *A* did was formulating hypotheses, making decisions about relevance of evidence to these hypotheses, about statistical independence between pieces of evidence given these hypotheses and, finally, about probability estimates. I claimed that all these decisions can only be reliably made by someone who is an expert in the domains of the various aspects of the case at hand. In the Breda Six this concerned, among other things, the time of rigor mortis, reliability of statements by the suspects and witnesses, information concerning prior convictions and prior criminal investigations, evidence of various traces like DNA, blood stains and hairs, statistical evidence concerning confession rates among various ethnic groups and various common-sense issues, such as the relevance of the fact that two of the six suspects worked in a snack-bar next door to the crime scene. In the Oosterland case the main evidence concerned statements of the suspects and witnesses, general knowledge about arson cases, information concerning prior convictions and prior criminal investigations and again various commonsense issues, such as how communities might turn against individuals and the relevance of friendships between suspects.

Let us now consider the case where *D* is the domain of Bayesian analysis of complex criminal cases, understood as comprising all the above issues. In my report, I formulated two general arguments against the truth of the first premise that A is an expert in this domain. First, expertise in the mathematics of Bayesian probability theory does not imply expertise in applying Bayes to a domain and, second, expertise in applying Bayesian probability theory in the domain of climate physics does not imply expertise in applying Bayes to the domain of complex criminal cases. The court in the Breda Six case instead ruled that A could be regarded as an expert for the purpose of the case. For space limitations an analysis of the court's justification of this decision has to be omitted. In the Oosterland case, the court did not discuss the issue of A's expertise but A himself discussed it in his written reply to my report. He admitted that he has no expertise in any of the relevant evidence domains of the case and argued that the value of his report did not lie in providing reliable posterior probabilities but in showing which questions had to be answered by the court. Against this I argued that even identifying the right questions in a complex criminal requires expertise in the relevant evidence domains.

Considering the critical questions of the scheme, personal bias (the second question) was not an issue. The first question (how credible is *E* as an expert source) is in fact a weaker version of the question whether the first premise (is *E* an expert in domain *D*?) is true: if the court in the Breda Six is followed in its decision that A can be regarded as an expert for the purpose of the case, then the arguments against this decision now become arguments that A's level of expertise is low. Such arguments are especially relevant when dealing with the third critical question (Is *P* consistent with what other experts assert?). In fact, A and I disagreed on a number of issues, so the court arguably had to assess the relative level of our respective expertise, and doing so is a kind of metalevel argumentation about the strength of arguments. Finally, the fourth question (Is *E*'s assertion of *P* based on evidence?) was used by me in forming arguments that most of A's probability estimates were not based on any data or scientific knowledge.

Concluding, [6]'s argument scheme from expert opinion is a good overall framework for analysing the debates about expertise in the two cases. On the other hand, most inter-

esting argumentation is not at the top level of this scheme but deeper down in the detailed arguments concerning the scheme's premises and critical questions.

4.3. Arguments from reasoning errors

In Section 4.2 I assumed that an expert asserts propositions but often an expert will assert an argument. Asserting an argument includes but goes beyond asserting its premises and conclusion: the expert also claims that the conclusion has to be accepted because of the premises. In many cases such an argument can be attacked by rebutting, undercutting or undermining it. However, sometimes a critic might want to say that the argument is inherently fallacious. This is not the same as stating an undercutting argument, since an undercutter merely claims that there is an exception to an otherwise acceptable inference rule. Especially in probabilistic and statistical reasoning real or claimed reasoning fallacies can be frequent, so arguments from reasoning errors deserve to be studied.

In the two cases of the present case study, several arguments about argument validity were exchanged. For reasons of space I can discuss just one example. In his report in the Oosterland case, A first estimated that the probability of fifteen arson cases in a town like Oosterland in a six-months period given the hypothesis that they were not related is at most one in a million. He then concluded from this that the fifteen arson cases considered by him cannot have been coincidence and that they must have been related. In my report I claimed that this argument is an instance of the prosecutor fallacy, since it confuses the probability that the fifteen incidents happen given that they are not related with the probability that the fifteen incidents are not related given that they happen.

One way to show that A's argument is fallacious is by giving a simple formal counterexample, for example, to specify for some E and H that $Pr(E \mid H) = Pr(E \mid \neg H) = 1/1.000.000$ so that the likelihood ratio of E with respect to H equals 1, so that the posterior probability $Pr(H \mid E)$ equals the prior probability $Pr(H)$, which can be any value.

From the point of view of argument visualisation one would like to have the following. For a given probabilistic statement φ, such as a link or probability in a Bayesian network, or a probability that is part of a likelihood ratio estimated by an expert, the user could click on the statement and be able to inspect the following argument:

Expert E asserts that $\psi_1, \ldots, \psi_n$
Expert E asserts that $\psi_1, \ldots, \psi_n$ imply φ
Therefore, φ because of $\psi_1, \ldots, \psi_n$.

Our example can be modelled with a combination of two applications of the expert testimony scheme combined with a deductive inference from their conclusions:

A is an expert on arson cases
E asserts that $Pr(incidents \mid \neg related) \leq 1/1.000.000$
E's assertion is within the domain of arson cases

$Pr(incidents \mid \neg related) \leq 1/1.000.000$

E is an expert in Bayesian reasoning
E asserts that P implies $Pr(related \mid incidents) >> 0.5$
E's assertion is within the domain of Bayesian reasoning

P implies $Pr(related \mid incidents) >> 0.5$

Here P is the conclusion of the first argument and $>>$ means 'much greater than'. The conclusions of these two arguments deductively imply $Pr(related \mid incidents) >> 0.5$.

My counterargument can be modelled as follows, where C stands for a description of the above-given counterexample:

$$\frac{\begin{array}{l} C \text{ implies that } P \text{ does not imply } Pr(related \mid incidents) >> 0.5 \\ C \end{array}}{P \text{ does not imply } Pr(related \mid incidents) >> 0.5}$$

In $ASPIC^+$ and similar formal argumentation systems this argument defeats the preceding one, since it is a deductive argument with universally true premises while its target is defeasible.

4.4. Analogical arguments

In the two case studies, several analogical arguments were used. The following version of the **argument scheme from analogy** is fairly standard; cf. [6, pp. 58,315].

$$\frac{\begin{array}{l} \text{Case } C_1 \text{ and } C_2 \text{ are similar in respects } R_1, \ldots, R_k \\ R_1, \ldots, R_n \text{ are relevant similarities as regards } P \\ P \text{ is true in case } C_1 \end{array}}{P \text{ is true in case } C_2}$$

Its two **critical questions** are: (1) Do cases C_1 and C_2 also have relevant differences? (2) Is Case C_2 relevantly similar to some other case C_3 in which P is false?

One use of analogy was in the Breda Six case, concerning the evidence that two of the three accused women worked in a snack-bar next door to the crime scene. In his report, A estimated the likelihood ratio of this "coincidence". A first estimated the denominator of this likelihood ratio (the probability of the coincidence given innocence of all six accused) as 1 in 500 (on grounds that are irrelevant here). He then estimated the numerator of this likelihood ratio (the probability of the coincidence given his guilt hypothesis) as 1, thus arriving at a strongly incriminating likelihood ratio of 500. Here he used an analogy with a hypothetical case in which a burglar breaks into a house by using a key of the house. Suppose a suspect is caught in possession of the key. According to A, possession of the key is a necessary element of the crime, so given guilt of the suspect the probability that he possesses the key is 1. In the same way, A argued, the coincidence in the Breda Six case is a necessary element in the crime, since A's guilt hypothesis was that at least some of the six accused were involved in the crime, *where one or more female accused lured the victim to the restaurant where the crime took place*. I criticised this on the grounds that, firstly, such luring can also be done by someone who does not work next door to the restaurant, such as the third female suspect; and, second, that the joint innocence of the two female suspects working next door to the restaurant is consistent with A's guilt hypothesis. So the coincidence cannot be regarded as a necessary element of the crime. I thus pointed at a relevant difference with A's hypothetical burglary case, in which possession of the key *is* a necessary element of the crime, thus using the first critical question of the analogy scheme.

4.5. Arguments from statistics

One might expect that in a probabilistic analysis of a complex criminal case, arguments from statistics to individual probability statements are frequent. Yet in my two cases most

probability estimates were not based on statistics; in just a few cases A used them to support his estimates. In some other cases A used a quasi-frequentist approach. For example, in the Oosterland case he estimated the probability that the suspect and someone else (a suspect in a related case) were best friends given the innocence hypothesis by first observing that Oosterland has 2400 inhabitants and then estimating that for men like the suspect there were 200 candidates in Oosterland for being his best friend, thus arriving at a probability of 1 in 200 given innocence of both. This illustrates that even if estimates are based on data, the step from data to probabilities can involve subjective assumptions (in this case that there were 200 candidates for being the suspect's best friend).

In its most basic form, **arguments from statistical frequencies** to an individual probability take the following form.

$$\frac{\begin{array}{l}\text{The proportion of } F\text{'s that are } G\text{'s is } n/m \\ a \text{ is an } F\end{array}}{Pr(Ga \mid Fa) \approx n/m}$$

This scheme is presumptive: there is no necessary relation between a frequency statement about a class and a conditional probability statement about a member of that class. Before considering the scheme's critical questions, let us look at how the first premise can be established. One way is by **statistical induction**:

$$\frac{\text{The proportion of investigated } F\text{'s that are } G\text{'s is } n/m}{\text{The proportion of } F\text{'s that are } G\text{'s is } n/m}$$

This scheme is not treated in the usual accounts of argument schemes, such as [6]. A full investigation of ways to criticise its use would lead us to the field of statistics, which is beyond the scope of this paper. For now it suffices to list two obvious **critical questions**: whether the sample of investigated F's is biased and whether it is large enough.

In my cases, A derived some statistical information from sources. For example, in the Breda Six case he used statistics reported in a criminological publication on the frequencies of confessions of denials among various ethnic groups in the Netherlands. The reasoning then becomes:

> E says that S is a relevant statistic, E is expert on this, therefore (presumably), S is a relevant statistic. Furthermore, S says that the proportion of investigated F's that were G's is n/m, therefore (presumably) the proportion of investigated F's that were G's is n/m.

The final conclusion then feeds into the scheme from statistical frequencies. In my report on the Breda Six case, I did not criticise A's specific selection of statistics on confessions and denials but I did note in general that selection of relevant and reliable statistics requires expertise in the subject matter at hand. I then observed that there was no evidence that A possessed relevant criminological expertise, thus in fact attacking the second premise of this line of reasoning. All this illustrates that even in reasoning from statistics the argument scheme from expert opinion is relevant.

I now turn to three possible **critical questions** of the scheme from statistical frequencies (there may be more).

1. *Is there conflicting frequency information about more specific classes?* This is the well-known issue of choosing the most specific reference class.

2. *Is there conflicting frequency information about overlapping classes?* This is a variant of the first question. If a belongs to two non-overlapping but non-inclusive classes F and H, then in general the proportion of F-and-H's that are G does not depend on the respective proportions of F's and H's that are G. So without further information nothing can be concluded on $Pr(Ga \mid Fa \wedge Ha)$.
3. *Are there other reasons not to apply the frequency?* For example, a might belong to some subclass for which commonsense or expert judgement yields different frequency estimates. For instance, in the Oosterland case, the probability estimated by A that the suspect and the other person were best friends given the innocence hypothesis ignored that both were outsiders in the community, that they had similar life styles and that one was previously convicted and the other was previously suspected of serial arson. Even if no statistics about these subclasses of adult male inhabitants of Oosterland exist, commonsense says that given these characteristics the probability of being best friends given innocence may be considerably higher than as estimated by A in his quasi-frequentist way.

Another scheme used by A in deriving probability estimates from statistics was the scheme from analogy. For example, in his report in the Oosterland case, A based his estimates of the probability of fifteen arson cases in a town like Oosterland in a half-year period given that no serial arsonist was active in Oosterland in that period among other things on statistics on arson in Japan and the United Kingdom. Applying this statistic to The Netherlands assumes that Japan and the United Kingdom are relevantly similar to the Netherlands as regards (serial) arson. This seems a quite common way of using statistics for deriving probability estimates. Here again the expertise issue comes up, since judging whether two countries are relevantly similar as regards (serial) arson requires domain expertise relevant to that question. Here too my general criticism was that there was no evidence that A, being a climate physicist, possesses such relevant expertise.

In sum, reasoning from statistics can be a combination of at least the following presumptive argument schemes: arguments from statistical frequencies, arguments from statistical induction, arguments from expert opinion and arguments from analogy.

5. Related research

One motivation underlying this paper is the design of support software for discussions about Bayesian analyses of complex criminal cases. In the medical domain, [7] present a similar system, which relates a medical BN to the clinical evidence on which it is based. Both supporting and conflicting evidence of a BN element can be represented in and shown by the system, as well as evidence related to excluded variables or relations. Three sources of evidence are modelled: publications, experts and data. Despite its argumentative flavour, the system is not based on an explicit argumentation model.

There is some earlier research on argumentation related to Bayesian modellings of criminal cases. [1] provide a translation from *ASPIC*$^+$-style arguments to constraints on Bayesian networks (BN). Their focus is different from the present paper in that their arguments are not about how to justify elements of BN but on incorporating the information expressed in an argument in the BN.

The closest to the present paper is [4], who proposes a set of source-based argument schemes for modelling the provenance of probability estimates in likelihood approaches.

Among other things, Keppens proposes schemes for expert opinion (a special case of the one in the present paper), for reasoning from data sets (not unlike the present scheme for reasoning from statistics) and for reasoning from generally accepted theories. In addition, Keppens proposes a set of schemes for relating source-based claims concerning the nature of subjective probability distributions (such as 'B has a [non-negative/non-positive] effect on the likelihood of C') to formal constraints on the probability distributions. Yet there is a difference in approach. Keppens primarily aims to build a formal and computational model, while this paper primarily aims to analyse how discussions about Bayesian modellings actually take place. Thus the present study complements Keppens' research. Also, the focus of Keppens' model is more limited than the present study in that it only models arguments about specific probability distributions.

6. Conclusion

In this paper two discussions about Bayesian modellings of complex criminal cases were analysed on their argumentation structure. Since this is a case study, the question arises how general the results are. It is hard to say to which extent the studied cases are typical, since Bayesian analyses of entire complex criminal cases are still rare in the courtroom. The usual uses of Bayes in the courtroom concern individual pieces of evidence, especially random match probabilities of forensic trace evidence (DNA, tyre marks, shoe prints, finger prints, glass pieces). Also, since I was involved in the two studied discussions, my analysis in the present paper may have been affected by a personal view. Nevertheless, with this in mind, the case study still warrants some preliminary conclusions. From a theoretical point of view the richness of argumentation about Bayesian modellings and the usefulness of several recognised argument schemes have been confirmed, two new argument schemes for interpretation arguments and arguments from statistics have been formulated, and a novel analysis of some subtleties concerning arguments from expert opinion has been given. From a practical point of view, the paper has identified a new use case for argumentation support tools, namely, support for argumentation about Bayesian probabilistic modellings of legal evidential reasoning.

References

[1] F.J. Bex and S. Renooij. From arguments to constraints on a Bayesian network. In P. Baroni, T.F. Gordon, T. Scheffler, and M. Stede, editors, *Computational Models of Argument. Proceedings of COMMA 2016*, pages 96–106. IOS Press, Amsterdam etc, 2016.

[2] I. Hacking. *An Introduction to Probability and Inductive Logic*. Cambridge University Press, Cambridge, 2001.

[3] H. Kaptein, H. Prakken, and B. Verheij, editors. *Legal Evidence and Proof: Statistics, Stories, Logic*. Ashgate Publishing, Farnham, 2009.

[4] J. Keppens. On modelling non-probabilistic uncertainty in the likelihood ratio approach to evidential reasoning. *Artificial Intelligence and Law*, 22:239–290, 2014.

[5] S. Modgil and H. Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument and Computation*, 5:31–62, 2014.

[6] D.N. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, Cambridge, 2008.

[7] B. Yet, Z.B. Perkins, N.R.M. Tai, and W.R. Marsh. Clinical evidence framework for Bayesian networks. *Knowledge and Information Systems*, 50:117–143, 2016.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-79

Noise Induced Hearing Loss: An Application of the Angelic Methodology

Latifa Al-Abdulkarim, Katie Atkinson, Trevor Bench-Capon,
Department of Computer Science, The University of Liverpool, UK

Stuart Whittle, Rob Williams and Catriona Wolfenden
Weightmans LLP, Liverpool, UK

Abstract. We describe the use of the ANGELIC methodology, developed to encapsulate knowledge of particular legal domains, to build a full scale practical application for internal use by a firm of legal practitioners. We describe the application, the sources used, the stages in development and the application. Some evaluation of the project and its potential for further development is given. The project represents an important step in demonstrating that academic research can prove useful to legal practitioners confronted by real legal tasks.

1. Introduction

Although AI and Law has produced much interesting research over the last three decades, [6], there has been disappointingly little take-up from legal practice. One important exception is the approach to moving from written regulations to an executable expert system based on the methods proposed in [9], which has been developed through a series of ever larger companies: Softlaw, Ruleburst, Haley Systems and, currently, Oracle[1], where it is known as *Oracle Policy Automation*. Key strengths of Softlaw and its successors were its well defined methodology, and its close integration with the working practices of its customer organisations. In the past year or so, however, there has been an unprecedented degree of interest in AI and its potential for supporting legal practice. There have been many articles in the legal trade press such as *Legal Business*[2] and *Legal Practice Management*[3]; UK national radio programmes such as *Law in Action*[4] and *Analysis*[5] and Professional Society events, such as panels run by the Law Society of England and

[1] http://www.oracle.com/technetwork/apps-tech/policy-automation/overview/index.html
[2] *AI and the law tools of tomorrow: A special report.* www.legalbusiness.co.ukindex.phpanalysis4874-ai-and-the-law-tools-of-tomorrow-a-special-report. All websites accessed in September 2017.
[3] *The Future has Landed.* www.legalsupportnetwork.co.uk. The article appeared in the March 2015 edition.
[4] *Artificial Intelligence and the Law.* www.bbc.co.uk/programmes/b07dlxmj.
[5] *When Robots Steal Our Jobs.* www.bbc.co.uk/programmes/b0540h85.

Wales[6]. At the ICAIL 2017 conference there was a very successful workshop on *AI in Legal Practice*[7]. The legal profession has never been so interested in, and receptive to, the possibilities of AI for application to their commercial activities. There are, therefore, opportunities which need to be taken. In this paper we describe the use of the ANGELIC (ADF for kNowledGe Encapsulation of Legal Information for Cases) methodology [2], developed to encapsulate knowledge of particular legal domains, to build a full scale practical application for internal use by a firm of legal practitioners, to enable mutual exploration of these opportunities.

In section 2 we provide an overview of the law firm for which the application was developed, the domain and the particular task in that domain at which the application was directed. Section 3 gives an overview of the ANGELIC methodology, while section 4 describes the sources used to develop the application. Section 5 discusses the process of capturing and refining the domain knowledge and section 6 the development of an interface to enable the knowledge to be deployed for the required task. Section 7 provides an evaluation of the project and section 8 concludes the paper.

2. Application Overview

The application was developed for Weightmans LLP, a national law firm with offices throughout the UK. Amongst other things, Weightmans act for employers and their insurance companies and advise them when they face claims from claimants for Noise Induced Hearing Loss (NIHL) where it is alleged the hearing loss is attributable to negligence on the part of the employer(s), or former employer(s), during the period of the claimant's employment. Weightmans advise whether the claimant has a good claim in law and, if appropriate, the likely amount of any settlement. Their role is thus to identify potential arguments which the employers or their insurance companies might use to defend or mitigate the claim. Compensators are thus looking to use the ADF primarily to improve how they can settle valid claims and pay proper and fair compensation in a timely manner when appropriate, whilst using the ADF to challenge cases which may have no basis in law or may be otherwise be defendable. In such an application it is essential that the *arguments* be identified. Black box pronouncements are of no use: it is the reasons that are needed. Note that the idea is to identify usable arguments: not to model any process of argumentation. The knowledge will here be deployed in a program not dissimilar from a "good old fashioned" expert system. This seems to meet the current task requirements, which are to support and so speed up decision making. The novelty resides in the methodology, which improves the elicitation process, and the form in which the knowledge is captured and recorded: unlike Softlaw it does not restrict itself to encoding written rules, but draws on other forms of documentation and expert knowledge, which may

[6]The full event of one such panel can be seen on youtube at www.youtube.com/watch?v=8jPB-4Y3jLg. Other youtube videos include Richard Susskind at www.youtube.com/watch?v=xs0iQSyBoDE and Karen Jacks at www.youtube.com/watch?v=v0B5UNWN-eY.

[7]https://nms.kcl.ac.uk/icail2017/ailp.php

include specific experiences such as previous dealings with a particular site and common sense knowledge, and structures this knowledge. In this way the ADF is not restricted to specific items of law, or to particular precedents, but can capture the wider negligence principles that experts distill from the most pertinent decisions. Since the knowledge encapsulated is a superset of what is produced in the CATO system [4], it could, were the task teaching law students to distinguish cases, equally well be deployed in that style of program. Note too that the analysis producing the knowledge, as in CATO, is performed by a human analyst and then applied to cases: the knowledge is not derived from the cases, nor is it a machine learning system.

3. Methodology Overview

The ANGELIC methodology builds on traditional AI and Law techniques for reasoning with cases in the manner of HYPO [5] and CATO [4] and draws on recent developments in argumentation, in particular Abstract Dialectical Frameworks (ADFs) [7] and ASPIC+ [11]. Formally ADFs form a three tuple: a set of nodes, a set of directed links joining pairs of nodes (a *parent* and its *children*), and a set of acceptance conditions. The nodes represent statements which, in this context relate to issues, intermediate factors and base level factors. The links show which nodes are used to determine the acceptability of other nodes, so that the acceptability of a parent node is determined by its children. The acceptance conditions for a node states how precisely its children relate to that node. In ANGELIC the acceptance conditions for non-leaf nodes are a set of individually sufficient and jointly necessary conditions for the parent to be accepted or rejected. For leaf nodes, acceptance and rejection is determined by the user, on the basis of the facts of the particular case being considered. Essentially the methodology generates an ADF, the nodes and links of which correspond to the factor hierarchy of CATO [4]. The acceptance condition for a node contain a prioritised set of sufficient conditions for acceptance and rejection and a default. Collectively, the acceptance conditions can form a knowledge base akin to that required by the ASPIC+ framework [10], but distributed into a number of tightly coherent and loosely coupled modules to conform with best software engineering practice [12]. Thus the acceptance conditions are used to generate arguments, and the ADF structure to guide their deployment.

The methodology is supported by tools [3] developed in parallel with, and informed by, this project to guide the knowledge acquisition, visualise the information, record information about the nodes such as provenance, and to generate a prototype to enable expert validation, and support refinement and enhancement. Once the knowledge is considered acceptable, a user interface is developed, in conjunction with those who will use the system in practice, to facilitate the input of the information needed for particular cases and present the results needed to support a particular task.

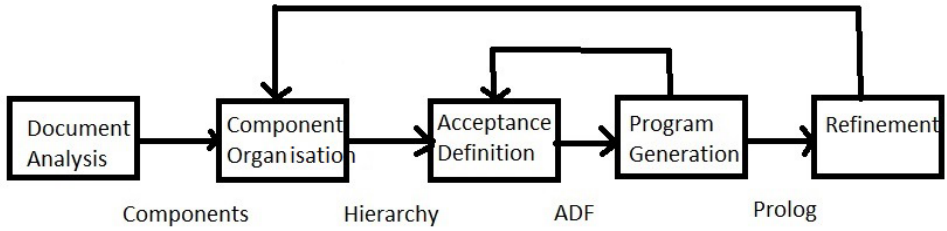


Figure 1. Knowledge Acquisition and Representation Stages

4. Sources

Several sources which were supplied by Weightmans were used to provide the knowledge of the Noise Induced Hearing Loss domain to which the ANGELIC methodology was applied.

- **Experts**: Weightmans made available domain experts to introduce the domain, provide specific documents and to comment on and discuss the developing representation.
- **Documents:** The documents included a 35 page information document produced by Weightmans for their clients, an 18 question check list produced by Weightmans to train and guide their employees and a number of anonymised example cases illustrating different aspects of the domain.
- **Users**: Potential users of the system were made available to assist in building and refining the interface.

Each of these sources played an invaluable role at various stages of the knowledge representation process, each making useful and complimentary contributions by providing different perspectives on the domain.

5. Representing the Knowledge

Following an introductory discussion of the domain, the workflow of the knowledge representation process comprises five steps, as shown in Figure 1:

1. Analyse the available documents and identify components and links between them.
2. Organise the components into an ADF.
3. Define the acceptance conditions for these components; the initial ADF is then reviewed by Weightmans and updated to accommodate the changes provided by the domain experts.
4. Extract a Prolog program from the acceptance conditions.
5. Run the program on the example cases to confirm the structure can generate the arguments in those cases and identify any necessary modification.

These stages are further described in the following sections.

5.1. Document Analysis

The initial discussion with the domain experts provided an excellent orientation in the domain and the key issues. These issues included the fact that claims were time limited, and so had to be made within 3 years of the claimant becoming aware of the hearing loss. Both actual awareness, usually the date of an examination, and constructive awareness (the date on which the claimant should have been aware that there was a problem) need to be considered. Then there is a question of the nature of the hearing loss: there are many reasons why hearing deteriorates, and only some of them can be attributed to exposure to noise. Then there is the possibility of contributory negligence: there is a Code of Practice with which the employers should have complied, and it is also possible that the employee was in part to blame, by not wearing the ear defenders provided, for example.

Next, the information document was used to identify the components that would appear in the ADF, putting some flesh on the skeleton that emerged from the initial discussion. The main document provides summaries of the main definitions, the development of the legal domain rules, the assessment of general damages for noise induced hearing loss cases, and Judicial College Guidelines for the assessment of general damages. Other medical conditions related to hearing loss are listed and described. At this stage components were identified, and where these were elaborated in terms of the conditions that were associated with them, links between these components could be identified. For example hearing loss can be sensioneural, but can also be attributed to a number of other factors: natural loss through aging, loss accompanying cardio-vascular problems, infections, certain drugs, etc. Only sensioneural loss can be noise induced, and so hearing loss arising from the other factors cannot be compensated. The document gives an indication of the various different kinds of hearing loss, and then further information of what may cause the various kinds of loss.

At the end of this phase we have a number of concepts, some of which are elaborated in terms of less abstract concepts, and some potential links. The next step is to organise these concepts in an appropriate structure.

The check list was kept back to be used after the concepts had been organised into a hierarchy, to determine whether the hierarchy bottomed out in sensible base-level factors. The check list comprised a set of 18 questions and a "traffic light" system indicating their effect on the claim. The idea was to associate base level factors with the answers to these questions. For example Question 1 asks whether the exposure ceased more than 3 years before the letter of claim: if it did not, the claim is *ipso facto* within limitation and other kinds of defence must be considered.

Similarly the cases were not used to build the initial ADF but were held back to provide a means of working through the ADF to check that the arguments deployed in those cases could be recovered from the ADF.

5.2. Component Organisation

The main goal now is to move from unstructured information gathered from the documentation to structured information. The main issues had been identified in

the initial discussion and the document analysis. These were used to identify and cluster the relevant intermediate predicates from the documents. These nodes were further expanded as necessary to produce further intermediate predicates and possible base level factors. The checklist was then used to identify, and where necessary add, base level factors. The documents from the sample of particular cases were used to provide examples of possible facts, and the effect these facts had on decisions. The result was a factor hierarchy diagram where the root shows the question to be answered, while the leaves show some facts from the sample cases. All this was recorded in a table that described the factors in the domain and their related children. For example, the Breach of Duty factor includes:

Factor: Breach of duty
Description: The employer did not follow the code of practice in some respect.
Children: Risk assessments were undertaken; employee was told of risks; methods to reduce noise were applied, protection zones were identified, there was health surveillance, training.

The children are the main things required of an employer under the code of practice, and so provide a list of the ways in which a breach of duty might have occurred. They may be further elaborated: for example noise reduction includes measures such as shielding the machinery and providing appropriate ear protection.

The final version of the ADF contains 3 issues, 20 intermediate nodes and 14 base level factors, with 39 links. For comparison, the ADF equivalent of CATO given in [2] contained 5 issues, 11 intermediate nodes, and 26 base level factors with 48 links. Thus CATO is larger, but NIHL has more internal structure. The nodes in the visual presentation of the ADF are annotated to show their provenance (the document and section in which they are defined or explained), and any of the checklist questions to which they relate.

5.3. Defining Acceptance Conditions

Once the nodes had been identified, acceptance conditions providing sufficient conditions for acceptance and rejection of the nodes in terms of their children were provided. These were then ordered by priority and a default provided. The particular cases were used to confirm that the arguments used in them could be recovered from the ADF. Continuing the Breach of Duty example:

Factor: Breach of duty
Acceptance conditions: Employee was not told of risks through the provision of education and training,
There were no measures taken to reduce noise,
Protection zones were not identified,
There was no health surveillance and no risk assessment.

Any of these are sufficient conditions to identify a breach of duty. If none of them apply to the case, we can assume, as a default, that there was no breach of duty, and so include rejection of the node as the default.

After this stage, the analysts and domain experts met to discuss and revise the initial ADF. Once a final ADF had been agreed, a Prolog program was produced from the acceptance conditions to suggest whether, given a set of facts, there might be a plausible defence against the claim.

5.4. Program Implementation

The program is implemented using Prolog. The program was created by ascending the ADF, rewriting the acceptance conditions as groups of Prolog clauses to determine the acceptability of each node in terms of its children. This required re-stating the tests using the appropriate syntax. Some reporting was added to indicate whether or not the node is satisfied, and through which condition. Also some control was added to call the procedure to determine the next node, and to maintain a list of accepted factors. We do not give the output here for reasons of space and commercial sensitivity, but its form is identical to that produced for Trade Secrets in [2]. The closeness between Prolog procedures and expressions of the acceptance conditions, each condition mapping to a clause within the Prolog procedure, makes the implementation quick, easy and transparent. The process of moving from acceptance conditions to Prolog code is essentially a mechanical rewriting into a template (supplying the reporting and control) and so is highly amenable to automation. Automated generation of the Prolog program from the ADF is planned as part of the development of the ANGELIC environment [3]. The program operates by:

- Instantiating the base level factors using the case facts;
- Working up the tree. Nodes are represented as heads of clauses, and each acceptance condition forms the body of a clause for the corresponding head, determining acceptance or rejection, with the set of clauses for the head completed by a default [8]. The program reports the status of the node and the particular condition which led to this status before moving to the next node.

The program provides a very transparent output that identifies precisely the path up the hierarchy and hence where any divergences from the expected outcomes occur. The program has been tested on a range of cases (additional to those originally supplied) identified by their base factors to evaluate the output and help the analysts and experts in detecting any errors or potential improvements.

5.5. Refinement

Both the initial ADF and the program were, again, shown to and discussed with the domain experts, who suggested corrections and enhancements. The corrections varied: some suggestions were made about considering missing information from the document, modifying the interpretation of existing acceptance conditions, or adding base factors or new parents to base facts. No changes were related to the main issues or intermediate predicates. As stated in [1], responding to changes in ANGELIC can be easily controlled since the changes affect nodes individually but, because of the modularisation achieved by the ADF, do not ramify through the rest of the structure. Refinement was an iterative process which was repeated until an ADF acceptable to the domain experts was obtained.

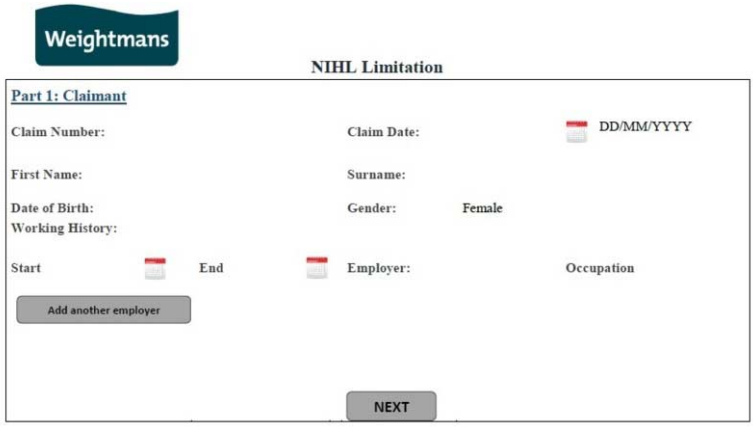


Figure 2. Screen for User Interface

6. User Interface

The ADF encapsulates knowledge of the domain, but this is required not for its own sake, but to add support in handling the analysis of the cases in the legal domain. To fulfill this task, a forms-based interface was designed in conjunction with some of the case handlers who carry out the task and so are the target users of the implemented system.

- The interface is designed to take as input the base level factors which correspond to the questions in the checklist used by the case handlers.
- These questions are organized in an order which makes good sense in terms of the task. First the questions related to the claimant's actual knowledge of the hearing loss are displayed. The answers to questions are used to limit the options provided in later questions and, where possible, to provide automatic answers to other questions.
- Three to four questions are used per screen to maintain simplicity.
- The input to the questions is as a drop down list with the given options (facts), or radio buttons when one option needs to be selected, or checklists for multiple options. Text boxes are also provided to input information particular to individual cases, such as names and dates, or to allow further information for some questions.
- All the questions must be answered, but default answers can be provided for some questions.

The designed interface enables the ADF to assist as a decision augmentation tool for the particular task. A sample screen is shown in Figure 2: the gender is pre-completed, but can be changed from a drop down menu.

7. Evaluation

Developing the application was intended to realise a number of goals, each offering a perspective for evaluation. Note that the system has not been fielded: it was

intended as a feasibility study and the programs are prototypes. Validation of its practical utility must await the fielding of a robust system engineered for operational use.

- The ANGELIC methodology had previously been applied only to academic examples. The desire here was to see whether it would also be effective when applied to a reasonably sized, independently specified, domain, intended to produce a system for practical use.
- Weightmans wished to come to a better understanding of the technology and what it could do for them and their clients.
- The methodology was designed to encapsulate knowledge of the domain using techniques representing the state of the art in computational argumentation. It was desired to see whether a domain encapsulated in this way could be the basis for a particular, practically useful, task in that domain.

Each of these produced encouraging results. The methodology proved to be applicable to the new domain without significant change, and could be used with the sources provided. Some desirable additional information that should be recorded about the nodes (such as provenance) was identified. The result was the specification and development of a set of tools to record and support the use of the methodology - the ANGELIC Environment [3].

Weightmans were encouraged that these techniques could prove useful to their business, and are currently exploring, with the University of Liverpool and others, options to take their investigations further.

From the academic standpoint, as well as confirming the usefulness and applicability of the ANGELIC methodology, the customisation for a particular task showed that the general knowledge encapsulated in the ADF can be deployed for a specific task by the addition of a suitable interface.

8. Concluding Remarks

The application of the ANGELIC methodology to a practical task enabled the academic partners in this project to demonstrate the utility of the methodology and identify possible extensions and improvements. The legal partners in the project were able to improve their understanding of the technology, what it could do for their business, and what development of an application would require of them. For the kind of application described here, the argumentation is all-important: the system is not meant to make a decision as to, or a prediction of, entitlement. Rather the case handlers are interested in whether there are any plausible arguments that could be advanced to challenge or mitigate the claim, or whether the arguments suggest that the claim should be accepted.

We believe that the success experienced for this task and domain is reproducible and look forward to using the methodology and supporting tools to build further applications, and to evaluating their practical utility when fielded. It should however, be recognised that the application developed here addresses only part, albeit a central part, of the pipeline. There is still a gap between the unstructured information which appears in a case file and the structured input necessary

to drive the program. In the above application this step relies on the skills of the case handlers, but there are other developments which could potentially provide support for this task, such as the tools developed by companies such as Kira Systems for contract analysis and lease abstraction[8]. It is to be hoped that this kind of machine learning tool might provide support for this aspect of the task in future. Similarly the interface is currently hand crafted and one off. It is likely that the process of developing a robust implementation from the animated specification provided by the Prolog program could benefit from tool support, such as that available from companies such as Neota Logic[9]. What has been described is essentially an exploratory study, but one which provides much encouragement and suggests directions for further exploration, and the promise that eventually robust decision support tools based on academic research will be used in practice.

References

[1] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. Accommodating change. *Artificial Intelligence and Law*, 24(4):1–19, 2016.

[2] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. A methodology for designing systems to reason with legal cases using Abstract Dialectical Frameworks. *Artificial Intelligence and Law*, 24(1):1–49, 2016.

[3] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. Angelic environment: Demonstration. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law*, pages 267–268. ACM, 2017.

[4] V Aleven. *Teaching case-based argumentation through a model and examples*. PhD thesis, University of Pittsburgh, 1997.

[5] K Ashley. *Modeling legal arguments: Reasoning with cases and hypotheticals*. MIT press, Cambridge, Mass., 1990.

[6] T Bench-Capon, M Araszkiewicz, K Ashley, K Atkinson, F Bex, F Borges, D Bourcier, P Bourgine, J Conrad, E Francesconi, et al. A history of ai and law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law*, 20(3):215–319, 2012.

[7] G Brewka, S Ellmauthaler, H Strass, J Wallner, and S Woltran. Abstract dialectical frameworks revisited. In *Proceedings of the Twenty-Third IJCAI*, pages 803–809. AAAI Press, 2013.

[8] K Clark. Negation as failure. In *Logic and data bases*, pages 293–322. Springer, 1978.

[9] P Johnson and D Mead. Legislative knowledge base systems for public administration: some practical issues. In *Proceedings of the 3rd ICAIL*, pages 108–117. ACM, 1991.

[10] S Modgil and H Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.

[11] H Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[12] R Pressman. *Software engineering: a practitioner's approach*. Palgrave Macmillan, 2005.

[8] see https://kirasystems.com/

[9] https://www.neotalogic.com/

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-89

Passing the Brazilian OAB Exam: Data Preparation and Some Experiments[1]

Pedro DELFINO [a,b] Bruno CUCONATO [b] Edward Hermann HAEUSLER [c]
Alexandre RADEMAKER [b,d]

[a] *FGV Direito Rio, Rio de Janeiro, Brazil*
[b] *Applied Mathematics School of FGV, Rio de Janeiro, Brazil*
[c] *Departamento de Informática, PUC-Rio, Brazil*
[d] *IBM Research, Brazil*

Abstract. In Brazil, all legal professionals must demonstrate their knowledge of the law and its application by passing the OAB exams, the national Bar exams. This article describes the construction of a new data set and some preliminary experiments on it, treating the problem of finding the justification for the answers to questions. The results provide a baseline performance measure against which to evaluate future improvements. We discuss the reasons to the poor performance and propose next steps.

Keywords. OAB, bar exam, question-answering, justification, logic

1. Introduction

The "Ordem dos Advogados do Brasil" (OAB) is the professional body of lawyers in Brazil. Among other responsibilities, the institution is responsible for the regulation of the legal profession in the Brazilian jurisdiction. One of the key ways of regulating the legal practice is through the "Exame Unificado da OAB" (Unified Bar Examination). Only those who have been approved on this exam are allowed to work as practising attorneys in the country. In this way, the it is similar to the US Bar Exam. Thus, the OAB exam provides an excellent benchmark for the performance of a system attempting to reason about the law.

This paper reports the construction of the data set and some preliminary experiments. We obtained the official data from previous exams and their answer keys from `http://oab.fgv.br/`. As our first contribution, we collected the PDF files, extracted and cleaned up the text from them producing machine-readable data (Section 2). [2]

An ideal legal question answering system would take a question in natural language and a corpus of all legal documents in a given jurisdiction, and would return both a correct answer and its legal foundation, i.e., which sections of which norms provide support

[1] The authors would like to thank João Alberto de Oliveira Lima for introducing us to the LexML resources, and Peter Bryant for his careful review of the article. The extend version of this article is available in `http://arademaker.github.io`

[2] All data files are freely available at `http://github.com/own-pt/oab-exams`.

for the answer. Since such a system is far from our current capabilities, we started with a simpler task. In [4] the authors report a textual entailment study on US Bar Exams. In the experiment, the authors treat the relationship between the question and the multiple-choice answers as a form of textual entailment. Answering a multiple choice legal exam is a more feasible challenge, although it is still a daunting project without restrictions on the input form, such as preprocessing natural language questions to make them more intelligible to the computer or restricting the legal domain. That is the reason we focused in the Ethics section of OAB Exams, one which is governed by only a few legal norms.

We have conducted three experiments in question answering (Section 3). To be able to provide the right justification for each question, we needed to have the text of the laws available. This is a particular challenge in the legal domain, as normative instruments are not readily available in a uniform format, suitable for being consumed by a computer program. Fortunately, using resources provided by the LexML Brasil project, we were able to collect and convert to XML format all the normative documents we needed (Section 2).

2. The data: OAB exams and norms

Before 2010, OAB exams were regional, only in 2010 were the exams nationally unified. In order to be approved, candidates need to be approved in two stages. The first phase consists of multiple choice questions and the second phase involves free-text questions. The first phase has 80 multiple choice questions and each question has 4 options. Candidates need at least a 50% performance.

Every year, there are 3 applications of the exam in the country. Concerning the exams statistics, the first phase is responsible for eliminating the majority of the candidates. Historically, the exam has a global 80% failure rate. Since 2012, the exams have revealed a pattern for which areas of Law the examination board focuses on and in which order the questions appear on the exam. Traditionally, the first 10 questions are about Ethics.

In the context of the OAB exam, Ethics means questions about the rights, the duties and the responsibilities of the lawyer. This is the simplest part of the exam with respect to the legal foundation of the questions. Almost all the questions on Ethics are based on the Brazilian Federal Law 8906 from 1994, which is a relatively short (89 articles) and well designed normative document. A minor part of the questions on Ethics is related to two other norms: (i) "Regulamento Geral da OAB" (OAB General Regulation, 169 articles) and (ii) "Código de Ética da OAB" (OAB ethics code, 66 articles). These two norms are neither legislative nor executive norms. Indeed, they are norms created by OAB itself. OAB's prerogative to do so is assured by the Law 8906.

We obtained the exams files in PDF format and we converted them to text using Apache Tika [3]. The final data comprises 22 exams totaling 1820 questions. A range of issues on the texts of the questions of the exams was identified. Many of the problems are similar to the ones found in the Bar Exams and described by [4]. For instance, some questions do not contain an introductory paragraph defining a context situation for the question. Instead of that, they have only meta comments, e.g. "assume that..." and "which of the following alternative is correct?". Some questions are in a negative form, asking the examinee to select the wrong option or providing a statement in the negative form

[3] https://tika.apache.org/.

such as "The collective security order **cannot** be filed by...". Moreover, some questions explicitly mention the law under consideration, others do not. Many questions present a sentence fragment and ask for the best complement among the alternatives, also exposed as incomplete sentences.

We sampled 30 questions on Ethics for analysis (from the 210 questions in all exams) and one of the authors manually identified the articles in the laws that justify the answer, creating our golden data set. The key finding was that, usually, one article on the Law 8906 was enough to justify the answer to the questions (15 questions). Less often, the justification was not in the Law 8906, but rather in OAB Regulation (3 questions), or on the OAB Ethics Code (8 questions). Three questions were justified by two articles in Law 8906, and another in jurisprudence from the Superior Court of Justice about an article from the Law 8906.

For the experiments, we also needed the norms in a format that preserved the original internal structure, i.e., the sections, articles, and paragraphs. The LexML [2] is a joint initiative of the Civil Law legal system countries seeking to establish open standards for the interchange, identification and structuring of legal information. The Brazilian LexML project has developed a XML schema called "LexML Brasil" and it maintains a public repository at `https://github.com/lexml` with one useful tool for our project, the parser of legal documents. The software receives as input a DOCX file and outputs it in XML file, according the LexML schema.

3. The Experiments

We borrow ideas from [7] to construct a similar experiments that run as follows: one collects the legal norms and preprocesses them performing tasks such as converting text to lower case, eliminating punctuation and numbers and, optionally, removing stop-words. After that, the articles of the norms are represented as TF-IDF vectors in a Vector Space Model (VSM) [1].

A base graph is then created, with a node for each article of a norm and no edges. When provided a question-answer pair, our system preprocesses the question statement and the alternatives in the same way as it does to the articles in the base graph. It turns them into TF-IDF vectors using IDF values from the document corpus.[4] The statement node is connected to every article node, and each article node is then connected to every alternative node, creating a connected digraph.

The edges are given weights whose value is the inverse cosine similarity between the connected nodes' TF-IDF vectors. The system then calculates the shortest path between question statement and answer item using Dijkstra's algorithm, and returns the article that connects them as the answer justification. The intuition behind such a method is that the more similar two nodes are, the lesser is the distance between them; as a document that answers a given query is presupposed similar to the question, it makes sense to retrieve the article in the shortest path between the statement and the alternative as a justification for the answer.

In our first experiment we had an ambitious objective: we had our system receive a question statement and its multiple alternatives, and we wanted it to retrieve the right

[4]This means that if a term occurs in the question statement or alternative but not on the legal norm corpus, its IDF value is 0.

answer along with its justification in the legal norm. When given the question and its alternatives, the system would add them to the base graph composed by the respective legal norm's articles. The system would return the shortest paths between the question statement and its alternatives, and the presumed justification would be the article connecting the statement and the closest alternative. The system's performance against this task was not impressive: although it chose the correct alternative 10 times, it only provided the correct justification for 8 of these.

Analyzing the system's output paints a more nuanced picture, however. In some cases, the system would find the correct justification article for the correct answer, but would pick as its putative answer another (incorrect) item, because it had a shorter path. Other times, it would not be capable of deciding between two (or more) answer items, as they all had a shortest path of the same distance. The following exam question is a sample case where this statistical approach to question answering is defective:

> The young adults Rodrigo (30-year-old), and Bibiana (35-year-old), who are properly enrolled in an OAB section [...] Considering the situation described, choose the correct alternative: A) Only **Bibiana** meets the eligibility criteria for the roles. B) Only **Rodrigo** meets the eligibility criteria for the roles. [...]

As one can see, these two options differ by only one word (the names of the fictional lawyers), and both are unlikely to be in the text of the legal norms, which means that they do not affect the calculation of similarity. A similar situation arises when one answer item makes a statement and another item denies this statement. In a question like this a system can only systematically report a correct answer if it has a higher-level understanding of the texts at hand: no bag-of-words model will suffice.

As our first experiment demonstrated that our simple system could not reliably pick the correct answer among four alternatives, we turned our attention to shallow question answering (SQA), where our system would only have to provide the correct legal basis for the answer provided along with the question. In our second experiment, we built separate base graphs from each of the three norms. For each question in our golden set, we added its statement and its correct answer to the base graph created from the norm which contains the article that justifies it. The sole task of our system, in this case, is to the determine which article from the provided norm justifies the answer. In this simpler form, performance was not bad: the system retrieved the correct article in 21 out of 30 question-answer pairs.

In our third experiment, we tried to see if our system could provide the correct article from the appropriate legal norm without us telling it which norm it should consider. Following this idea, we have taken the articles from all norms and built a single base graph. For each question in our golden set, we again added its statement and correct answer as nodes connected to all article nodes in the graph, and then calculated the shortest path between them to retrieve the system's putative justification to the question-answer pair. The system now had to retrieve the correct article among articles from all norms – which, being in the same legal domain, had similar wordings and topics – therefore increasing the difficulty of the task. Despite this, its performance did not plunge: it scored the right article in 18 out of the 30 question-answer pairs.

4. A possible logic-based approach

One of the key observations that emerge from the results in Section 3 is the importance of logical reasoning for our final goal of constructing a system to pass the OAB exam with a full understanding of the questions and laws. For the future, we aim to investigate how to enrich the data with lexical information and syntactic dependencies as an intermediary step toward a semantic representation of the questions and laws statements. Nevertheless, we have to decide what should be an adequate logic language to represent laws and the deep semantics from the text statements. Since the adequacy of a logic language can be evaluated even before a procedure to obtain logic expressions from natural language texts is developed, we present some preliminary discussion about one possible logic.

In [5] we discuss how Kelsen's [6] pure theory of law points out a framework that takes into account the legal knowledge forming a collection of individual, legally valid statements. Thus, each legally valid statement may be seen as an inhabitant among the many individual laws of the represented legal system. The natural precedence existing between individual legal statements can be taken as a pre-order relation on the legal statements. The legal principle that rules the stability of the law implies that the precedence of individual laws preserves properties (decisions, conditions of applicability, adequate fora, etc) regarding them. In the presence of this natural precedence order between legally valid statements, the intuitionistic interpretation of subsumption between concepts A and B ($A \sqsubseteq B$) reflects more adequately the structure of existing legal systems than its classical interpretation counterpart.

To illustrate the use of *i*ALC for reasoning over the OAB exams questions, let us consider the translated question and its correct alternative:

> Three friends graduated in a Law School in the same class: Luana, Leonardo, and Bruno. Luana, 35 years old, was already a manager in a bank when she graduated. Leonardo, 30 years, is mayor of the municipality of Pontal. Bruno, 28 years old, is a military policeman in the same municipality. The three want to practice law in the private sector. Considering the incompatibilities and impediments to practice, please select the correct answer. [...] C) The three graduates, Luana, Leonardo, and Bruno, have functions incompatible with legal practice. They are therefore prohibited from exercising private practice. (CORRECT) [...]

The justification of the answer to this question is obtained in the Law 8906, article 28. [5] The relevant fragments of this article, translated into English, are:

> Legal practice is incompatible, even for self-defense, with the following activities: I - head of the Executive and members of the Bureau of the Legislative Branch and their legal substitutes; [...] V - occupants of positions or functions linked directly or indirectly the police activity of any nature; [...] VIII - occupiers of management positions in financial institutions, including private ones. [...]

In *i*ALC, the Law 8906 is formalized as a concept defined as the intersection of the concepts from its articles, that is, $Law_{8906} \equiv Art_1 \sqcap \ldots \sqcap Art_{28} \sqcap \ldots \sqcap Art_{87}$. Article 28 in turn is also further formalized as the intersection of the concepts from its paragraphs, $Art_{28} \equiv P_1 \sqcap P_2 \ldots$. The paragraph VIII is formalized by the two

[5] The complete text can be found at `http://bit.ly/29gZc83`

concepts $Lawyer \sqsubseteq \neg Financial$ and $Financial \sqsubseteq \neg Lawyer$. Paragraph V is formalized by $Lawyer \sqsubseteq \neg Police$ and $Police \sqsubseteq Lawyer$. Finally, paragraph I by $Lawyer \sqsubseteq \neg ChiefCouncil$ and $ChiefCouncil \sqsubseteq \neg Lawyer$. The *Lawyer* concept can be read as the set of valid legal statements (VLS) about lawyers. That is, each concept can be thought as the set of VLSs where it *holds*. From the statements of the question, we have the hypotheses $lual : Laywer$ (Luana acts as lawyer), $leoal : Lawyer$ and $bal : Lawyer$. Using the deductive system for *i*ALC [5], we can prove that Luana, Bruno and Leonardo can not act as lawyers.

$$\dfrac{\dfrac{lual : Police \qquad Police \sqsubseteq \neg Lawyer}{lual : \neg Lawyer} \qquad [lual : Lawyer]}{\dfrac{lual : \bot}{\neg(lual : Lawyer)}}$$

5. Conclusion and Future Works

We presented a new data set with all Brazilian OAB Exams and their answer keys jointly with three Brazilian norms in LexML format. Furthermore, we also presented some preliminary experiments with the goal of constructing a system to pass in the OAB exam. We obtained reasonable results considering the simplicity of the methods employed. For the next steps, we can construct the TF-IDF vectors using lemmas of the words, possibly increasing the similarities. We can also add edges between articles, considering that 10% of our golden set includes more than one article as justification. We also plan to use the OpenWordnet-PT [3], properly expanded with terms of the legal domain.

Finally, the results of the experiments presented here clearly show that we need 'deep' linguistic processing to capture the meaning of natural language utterances in representations suitable for performing inferences. That will require the use of a combination of linguistic and statistical processing methods. The final objective is to obtain formal representations, encoded in *i*ALC or another variant, from the texts ready for formal reasoning.

References

[1] D Manning Christopher, Raghavan Prabhakar, and Schütze Hinrich. Introduction to information retrieval. *An Introduction To Information Retrieval*, 151:177, 2008.

[2] João Alberto de Oliveira Lima and Fernando Ciciliati. LexML brasil: versão 1.0. available at `http://projeto.lexml.gov.br/`, December 2008.

[3] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper), 2012.

[4] Biralatei Fawei, Adam Z Wyner, and Jeff Pan. Passing a USA national bar exam: a first corpus for experimentation. In *Language Resources and Evaluation*, pages 3373–3378, 2016.

[5] Edward Hermann Haeusler, Valéria de Paiva, and Alexandre Rademaker. Intuitionistic logic and legal ontologies. In *Proc. JURIX 2010*, pages 155–158. IOS Press, 2010.

[6] Hans Kelsen. *General theory of norms*. Oxford Univ. Press, USA, 1991.

[7] Alfredo Monroy, Hiram Calvo, and Alexander Gelbukh. Using graphs for shallow question answering on legal documents. In *MICAI 2008: Advances in Artificial Intelligence: 7th Mexican International Conference on Artificial Intelligence, Atizapán de Zaragoza, Mexico, October 27-31, 2008 Proceedings*, pages 165–173. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-95

Answering Legal Research Questions About Dutch Case Law with Network Analysis and Visualization

Dafne VAN KUPPEVELT [a], Gijs VAN DIJCK [b]
[a] *Netherlands eScience Center*
[b] *Maastricht University*

Abstract. The availability of large collections of digitalized legal texts raises an opportunity for new methodologies in legal scholarship. Analysis of citation networks of case law gives insight into which cases are related and to determine their relevance. Software tools that provide an graphical interface to case law networks are required in order to enable non-technical researches to use network analysis methodologies. In this study, we present open source software for the analysis and visualization of networks of Dutch case law, aimed for use by legal scholars. This technology assists in answering legal research questions, including determining relevant precedents, comparing the precedents with those identified in the literature, and determining clusters of related cases. The technology was used to analyze a network of cases related to employer liability.

Keywords. network analysis, case law, visual analytics

1. Introduction

Legal documents such as case law and legislation are increasingly made available to the public. In the Netherlands, the government provides a dataset with the most important case law as XML files on *www.rechtspraak.nl.* The LiDO data bank[1] (also provided by the Dutch government) offers a linked data platform linking different legal sources. It contains meta data of Dutch and European case law and legislation and, more recently, also computer-identified references in Dutch case law to legislation and other case law. Computer-processing of these datasets allows legal researchers to investigate a large number of cases, in contrast to traditional methods that focus on a few, allegedly relevant cases. One way to represent a collection of data with references is as a network. Since decisions of judged can form precedents for future cases, the references between cases represent how case law is made by judges. Thus a network representation reveals the structure in the data and provides insight into legal questions that are very difficult to answer by looking at cases individually. However, graphical interfaces to the underlying data structures are needed for non-technical legal scholars.

[1] http://linkeddata.overheid.nl/front/portal/lido

Several previous studies have applied network analysis on case law. Fowler et al. [1] apply network analysis on Supreme Court cases in the US to determine relevance based on network statistics. Winkels [2] applies network statistics on Dutch case law, Derlén and Lindholm [3] on European case law and Schaper [4] on European direct tax law. A few web-based legal network visualization exist, such as EUcaseNet [5] and LexMex [2]. There are also commercial tools that apply network analysis and visualization for legal practice, such as Ravellaw [3] for US case law, and Juribot [4] for Dutch legal data. Despite many experiments with network analysis on legal data, generic tools that empower a non-technical legal researcher to explore the structure of legal data are difficult to find. This is most likely due to the relative lack of publicly available APIs to several sources of law. Consequently, collecting data for network analysis is often time-consuming and requires technical expertise to transform the data into the proper format for visualization and subsequent analysis.

Our work takes the viewpoint of the legal scholar without a technical background, and provides a generic open source technology based on publicly available data. We evaluate the technology from the legal perspective, by studying an example network with the technology and showing how legal research question can be answered for this network.

2. Network Analysis on Case Law

The network approach views cases as *nodes* in the network, and references between cases as *edges*. Following previous related studies, the following network statistics have been defined as possibly meaningful for research in case law networks:

- *In-degree*: the number of incoming references. Considering the concept of precedent in case law, cases that are referenced frequently, are more likely to be important than cases that are not frequently cited.
- *Out-degree*: the number of outgoing references. Cases with a large out-degree can be considered well-grounded, since the decision is based on many sources [3].
- *HITS hubs and authorities*: the HITS algorithm [6] gives an authority score, meaning how much a node is cited by nodes that are 'hubs', and a hub score, meaning how much a node is cited by nodes that are 'authorities'. Nodes with a high hub-score thus represent cases that have many citations to authoritative cases.
- *Relative in-degree*: the number of incoming references, corrected for the number of cases that exist later in time. Introduced by Tarissan and Nollez-Goldbach [7], this metric attempts to account for the fact that early cases have a larger in-degree, simply because there is more opportunity to be cited by succeeding cases.
- *Betweenness centrality*: A measure for how many shortest paths go through a node, i.e. how important the node is to connect the network. A large betweenness centrality can indicate that a case connects several subareas of the network.
- *Pagerank*: The PageRank algorithm [8] assigns scores to nodes based on the scores of incoming references. Although it has been argued that PageRank is difficult to use for case law networks [2], Derlén and Lindholm [3] have used it to determine importance of cases.

[2] http://www.lexmex.fr/
[3] http://ravellaw.com/
[4] https://referenties.semlab.nl/

Additionally, the Louvain community detection method [9] is used to identify clusters of nodes that form a community in the network. This can be used in the visualization to color the nodes by community, so that the parts of the network are visually separated. Tarissan et. al. [10] have shown that applying network statistics on a subnetwork around a certain topic can highlight landmark cases, in contrast to analysis on a complete network of a specific court. Therefore, we aim to develop an application for researchers to collect networks based on a specific set of cases related to a topic of interest.

3. Technology

The technology developed in this project consists of two web-based tools: the *caselawnet querier* [11] and the *case-law-app visualization* [12]. The querier application was built on top of the search API of rechtspraak.nl and used the link extractor API of LiDO. The querier allows users to search case law on keywords, and construct networks from a collection of cases. It is possible to include cases that were not in the original search result, but that are linked to one of the cases in the result. It is also possible to construct a network based on a user-defined set of cases, optionally including linked cases. The caselawnet querier is built in Python and Flask[5], and uses NetworkX[6] for calculating network statistics. The networks can be downloaded as csv-files of the nodes or links for further data analysis, or as JSON files to use in the visualization application.

Our visualization tool uses the Javascript library SigmaJS for graph rendering. The ForceAtlas2 layout algorithm [13] is used to position the nodes. In this algorithm, nodes that are connected through an edge are pulled closer together, resulting in a layout that emphasizes the structure of the network. The user can filter or change the appearance of the network, based on attributes of the nodes. These attributes include metadata of the cases, such as the court or year of the decision. It also includes network statistics, as described in the previous section.

4. Results in Legal Research

We discuss a number of research questions in legal research that can be answered with the technology presented in this paper. We illustrate these with an example network consisting of a set of 154 cases of Dutch supreme court, related to employer liability. This dataset was collected manually between 15 January and 5 April 2016, when computer-identified references were not available yet. Cases that were not directly about employer liability, but that were referenced by one of the employer liability cases were included as well. The citation network was enriched with meta data and network statistics using *caselawnet* and visualized with the *case-law-app*.

Which thematic subareas exist in the collection of cases? The Louvain method provides a starting point to define thematic subareas. In the employer liability network, 28 communities are detected by the Louvain method. Selecting only communities that are connected to the rest of the network, six communities are left. A qualitative analysis of

[5] http://flask.pocoo.org/

[6] https://networkx.github.io/

Table 1. Most cited decisions in the causality clusters

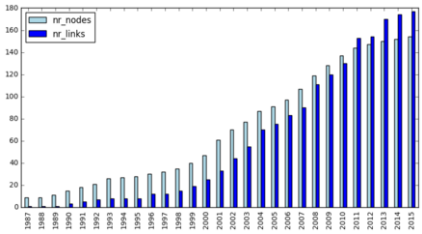
Name	ECLI	In-Degree	Authority	Relative In-Degree
Unilever/Dikmans	ECLI:NL:HR:2000:AA8369	6	0.165	0.055
Havermans/Luyckx	ECLI:NL:HR:2006:AW6166	4	0.138	0.069
Nefalit/Karamus	ECLI:NL:HR:2006:AU6092	3	0.087	0.051

the legal content of the cases inside the communities shows thematic coherence within the communities. The themes of the six connected clusters are identified as follows: 1. Asbestos; 2. Duty of care (extent), recipients' liability, contributory negligence, gross negligence; 3. Causality; 4. Losses while carrying out work, duty to insure; 5. Reversal rule; 6. Experts, evidentiary burden. These themes partly overlap with the main subjects of employer liability mentioned in literature. For example, the reference work Asser [14] describes 22 subjects in the area of employer liability, among which are the subjects *duty of care*, *duty to insure*, *reversal rule*. Since most legal literature mentions many different subjects and doesn't attribute a subject to every case, it is difficult to make a quantitative comparison with the subjects from the clusters.

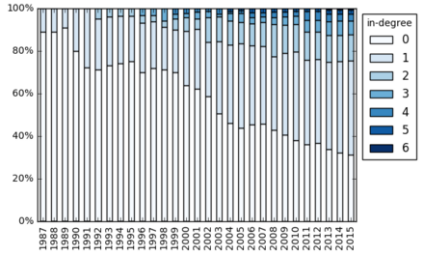
What are cases that have an important precedent value in the collection of case law? Network statistics can be used to identify important cases based on high relevance scores, both for the complete network and for each of the thematic clusters. The metrics in-degree, authority and relative in-degree all give an indication of the precedent value of a node in the network. We will thus study cases with high values for one or more of these measures. For the employer liability network, we will give an example for one of the *causality* cluster, but the same can be done for the rest of the network. The three most cited cases in the cluster are shown in Table 1. *Unilever/Dikmans* has the highest in-degree, but *Havermans/Luyckx* has the highest relative in-degree. Looking at the content, *Unilever/Dikmans* decides that the reversal rule is applicable, which forms an important precedent. *Havermans/Luyckx* builds on this decision by adding that it is the employee who needs to argue convincingly that he suffers from health problems that may be caused by the exposure to health hazards. In this way, *Havermans/Luyckx* takes over the role of precedent for cases that determine when the burden of evidence regarding causality is shifted from the employee to the employer. We also look at the case with the highest *betweenness centrality*, which is *Fransen/Stichting Pasteurziekenhuis* (ECLI:NL:HR:1999:AA3837). This case indeed connects clusters about asbestos, causality, the duty to insure and the reversal rule. The decision was a landmark case after the introduction of the 1992 Dutch Civil Code that gave direction as to how to interpret and apply the then new provision on employer liability.

How does the importance of cases change over time? The way in which judges apply the law depends on the context of the decision, such as the time period in which decision was taken. The network structure changes over time, and thus the network statistics such as in-degree and relative in-degree are not static. This can be explored by plotting the variables over time. The plots were created with Python using the *caselawnet* software. The code is available as iPython Notebook[7]. The size of the network (number of nodes and number of links) over time is plotted in Figure 1a. Naturally, the number of cases grows over time, as well as the number of links. The number of links grows faster than the number of nodes, which means that the network becomes 'denser' over time. By

[7] `https://github.com/caselawanalytics/CaseLawAnalytics/blob/master/notebooks/TimeAnalysis.ipynb`



(a) The size of the network over time.



(b) The distribution of in-degree over the years: the fraction of cases for each value of in-degree are shown.

Figure 1. The development of the network over time

looking at the distribution of in-degree varying over the years, as shown in Figure 1b, we see the distribution shifting so that relatively fewer cases have no incoming references. This indicates that over time, when the network becomes denser, the new links cite more *different* cases, instead of the cases that already had a large in-degree. Note that the shift might be partly caused by the way the network was constructed: cases that were referenced by one of the employer liability cases were included as well, so the selection of cases is based on the references in the most recent network. A solution for this would be to construct the networks in the exact same manner for each year, which for this network would require manual work that was outside of the scope of this paper.

5. Discussion and Conclusion

We have shown that the *caselawnet* and *caselawapp* prototype applications enable legal scholars to conduct empirical research using large number of decisions, by providing a graphical interface to the Dutch *rechtspraak.nl* and *LiDO* data sources. We used the technology for the analysis of a network about employer liability in Dutch Case law, which resulted in the identification of thematic sub-areas and important precedent cases in these sub-areas. It also provided insight in the evolution of the legal network over time.

Network visualization can assist in learning about a specific area of law, not only in research but also in education and legal practice. To familiarize with a new legal area, the tool can be used to answer questions about the data, such as: What are the most-cited cases about? What thematic subareas exist in this collection? In what time period do many (important) cases appear? Do different courts, in particular the Supreme court, refer to different cases? How dense and how connected is the network? In other words, how many citations do cases have on average and how much interaction is there between different subareas?

Since network analysis is a novel methodology for legal researchers, guidance and education is required to get the legal community familiarized with network analysis. The network statistics, other than in-degree and out-degree, have no simple meaning and require thorough interpretation. Users have to be careful not to make false assumptions based on the technology, such as spatial closeness in the ForceAtlas2 layout. In addition, the manner in which the network was constructed can influence the results of the network

analysis, as we have shown in the analysis of the distribution of in-degree over time. It is therefore important that the network construction is done carefully and documented well.

The data collection tool should be further improved, by combining keyword search with network exploration and filter options to assemble data collections in an interactive manner, possibly also including European legal sources (HUDOC and EurLex). Natural Language Processing techniques could assist in identifying themes of communities in the network. Research about developments of case law networks over time, as presented in this paper, could benefit from further visualizations that show the temporal aspect. Examples of such visualizations are animations of the network changing over time and interactive graphs that show statistics such as network size and in-degree varying in time.

The methodology can be evaluated further in a more quantitative manner, by asking a group of subjects to perform a series of tasks using the software and measuring their time use and satisfaction. The Louvain method could be further evaluated by manually annotating a data set with themes and validating the communities against the annotations. Lastly, it would be valuable to apply the network methodology to many different legal domains and compare the analyses across the networks. The methodology will eventually prove most value if it leads to novel research results in different legal domains.

References

[1] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck, "Network analysis and the law: Measuring the legal importance of precedents at the US supreme court," *Political Analysis*, vol. 15, no. 3, pp. 324–346, 2007.

[2] R. Winkels, J. d. Ruyter, and H. Kroese, "Determining authority of Dutch case law," *Legal Knowledge and Information Systems*, vol. 235, pp. 103–112, 2011.

[3] M. Derlén and J. Lindholm, "Goodbye van Gend en Loos, Hello Bosman? using network analysis to measure the importance of individual cjeu judgments," *European Law Journal*, vol. 20, no. 5, pp. 667–687, 2014.

[4] M. G. Schaper, "A computational legal analysis of Acte clair rules of eu law in the field of direct taxes," *World Tax Journal*, vol. 6, no. 1, 2014.

[5] N. Lettieri, A. Altamura, A. Faggiano, and D. Malandrino, "A computational approach for the experimental study of EU case law: analysis and implementation," vol. 6, 12 2016.

[6] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[7] F. Tarissan and R. Nollez-Goldbach, "Temporal properties of legal decision networks: a case study from the international criminal court," in *28th International Conference on Legal Knowledge and Information Systems (JURIX'15)*, 2015.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[10] F. Tarissan, Y. Panagis, and U. Šadl, "Selecting the cases that defined Europe: complementary metrics for a network analysis," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 661–668.

[11] D. van Kuppevelt, "caselawanalytics/caselawanalytics," Aug. 2017, doi: 10.5281/zenodo.854541. [Online]. Available: https://github.com/caselawanalytics/CaseLawAnalytics

[12] ——, "Nlesc/case-law-app," Jul. 2017, doi: 10.5281/zenodo.596839. [Online]. Available: https://github.com/NLeSC/case-law-app

[13] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software," *PLOS ONE*, vol. 9, no. 6, pp. 1–12, 06 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0098679

[14] G. H. van Voss, *Asser 7-V Arbeidsovereenkomst*. Kluwer, 2012.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-101

On Annotation of the Textual Contents of Scottish Legal Instruments

Adam WYNER [a,1] Fraser GOUGH [c], Francois LEVY [b], Matt LYNCH [c], and Adeline NAZARENKO [b]
[a] *University of Aberdeen, Aberdeen, United Kingdom*
[b] *LIPN, Paris 13 University – Sorbonne Paris Cité & CNRS, Paris, France*
[c] *Parliamentary Counsel Office, Scottish Government, Edinburgh, United Kingdom*

Abstract. LegalRuleML is a developing standard for representing the fine-grained semantic contents of legal texts. Such a representation would be highly useful for Semantic Web applications, but deriving formal rules from the textual source is problematic; there is currently little in the way of methodology to systematically transform language to LegalRuleML. To address this, we outline the purposes, processes, and outputs of a pilot study on the annotation of the contents of Scottish legal instruments, using key LegalRuleML elements as annotations. The resulting annotated corpus is assessed in terms of how well it answers the users' queries.

Keywords. semantic annotation, legal text processing, markup language, methodology

1. Introduction

There is an increasing demand for tools enabling fine-grained semantic access to legal sources, that is, for search tools that go beyond keyword search [2], such as Semantic Web applications to link, search, extract, and draw inferences with respect to the contents of and relations amongst legal rules. This paper presents a pilot study that shows how some of these demands, *e.g.* for search and extraction, can be addressed. However, formalizing the rule information present in legal sources is a complex task that cannot be automated due to the complexities of legal language and information. Nonetheless, some progress can be made to annotate the semantic structure of the source texts as well as to comply with existing documents and legal rule standards to ensure interoperability and linkability. The challenges are twofold: to make annotations that address users' interests; to make the annotation task feasible for legal people and in a form amenable to incremental refinement. In the experiment reported, a small corpus of legal instruments is translated to LegalRuleML, an XML mark-up language for legal rules [1]. By way of evaluation, we used the sample questions provided by the use case partners to query the annotated corpus; the results demonstrate the utility of the approach.

In the following, we present the use case requirements (Sec. 2), the annotations and corpus (Sec. 3), the methodology and tools (Sec. 4) and our preliminary outputs (Sec. 5).

[1]Corresponding Author: azwyner@abdn.ac.uk. We thank the funding from the University of Aberdeen's Impact, Knowledge Exchange, and Commercialisation Award for this 10 week study. This work was also supported by the French National Research Agency (ANR-10-LABX-0083) in the context of the Labex EFL. We also thank the student staff: A. Andonov, A. Faulds, E. Onwa, L. Schelling, R. Stoyanov, and O. Toloch.

2. Requirements

We started with the requirements set by the parliamentary counsel of the Scottish Government's Parliamentary Counsel Office, which is working to improve internal legislative drafting and information services and to provide legislative information "as a platform" for a robust ecosystem of legal services. A key part of this effort is to provide a corpus of law in electronically readable form which can be queried.

We were provided with questions to answer:

1. What are all the offences and associated penalties or defences?
2. What prohibitions apply to tobacco products?
3. What obligations have been placed on which entities, *e.g.* shop owners?
4. What permissions are given to Scottish Ministers?
5. Given a provision, what are related overriding or reparation provisions?

Answering such questions requires a substantial semantic analysis of the text. The challenge is to develop a level of analysis and XML representation which satisfies the questions. A sound methodology of annotation is necessary to get a corpus that can be further used as a gold-standard for evaluation and machine learning.

3. LegalRuleML and Annotations

Large corpora of legal texts must be machine-readable [2]. XML standards have been developed for document structure (Akoma Ntoso[2]) and semantic content (LegalRuleML [1]). Complying with such standards allows materials to be amenable to Semantic Web technologies. Yet analysing the semantic content of legal documents in terms of XML is particularly daunting given the nature of linguistic representation; there is a significant gap between the linguistic and formal representations of the law.

LegalRuleML is a proposed OASIS standard for rich XML representation, which has elements to represent legal content. In addition, it adopts a restricted set of XML elements from RuleML, a markup language for predicate logic rules[3]. In order to develop the means to translate from natural language to LegalRuleML, it has been argued that some intermediate annotation language is essential to get a "first draft" of the contents of the legal text as well as to help address linguistic ambiguities and interpretive issues [5]. In this project, we only used a small palette of LegalRuleML elements which associate with text annotations:

- Permission: the bearer is allowed to do something or be in a state.
- Obligation: the bearer is bound to do something or be in a state, for otherwise, the bearer is in violation.
- Prohibition: the bearer is bound not to do something or be in a state, for otherwise, the bearer is in violation.
- Constitutive: a definition.
- Override: an indication that one legal rule takes precedence over another.
- Reparation: an indication of a link between a penalty and a prescriptive norm.
- Penalty: a sanction.

[2] http://www.akomantoso.org/
[3] http://ruleml.org/index.html

1 Prohibition of tobacco displays etc.

(1) [prohibition 1 A person who in the course of business displays or causes to be displayed tobacco products or smoking related products in a place where tobacco products are offered for sale commits an offence prohibition]

Figure 1. Annotations on text

This small, coarse-grained palette of LegalRuleML elements was useful in addressing some key initial issues. Given an iterative, extensible development process, we can work with other elements in later phases. Similarly and for our purposes here, we do not work with document structure, which would be annotated in Akoma Ntoso, though in future iterations, such information will be important.

While LegalRuleML is explicit, application of the elements to text is not transparent. That is, the list of elements and their definitions are not sufficient for the consistent and accurate application of the annotations to text, nor is there clarification about how to analyse source text into LegalRuleML. Thus, an annotation methodology is required to connect text to LegalRuleML.

4. Methodology, Corpus, and Tools

To use LegalRuleML elements for annotation, we "hide" the technical structure of LegalRuleML from legal annotators, whose task was to understand the content. We provided annotators with a simplified set of annotations, where the relevant sentences are bracketed, labelled/typed, and possibly related via indices (see Figure 1). It is important to emphasise that we have "repurposed" LegalRuleML elements as labels/types for text annotation in order to associate text annotations with LegalRuleML representations; we have not thereby created an auxiliary markup language. On the semantic side, we developed guidelines with illustrations of regular and irregular examples to help tackle semantic issues. Adjudication and revision (of annotations and/or the guidelines) were essential.

The project employed four annotators for six weeks; they were students from different disciplines, but with some legal and linguistic training. Each original document was annotated by two legal annotators, who reviewed and commented on one another's work. Three "meta" annotators adjudicated the annotated documents. Once adjudicated, the resulting documents were translated into valid LegalRuleML files by LegalRuleML analysts. The annotators used an annotation manual, which was developed to guide annotations. During the annotation process, comments were added to the document, facilitating and tracking discussion. We reported the main issues and ambiguities in the manual.

For a corpus of texts, we have 10 legal instruments provided by the Scottish Government's Parliamentary Counsel Office (41,859 words, $\sim$ 140 pages)[4]. All bear on Scottish smoking legislation and regulation. The average word count per document is 4185.9, with a maximum word count of 12739 and a minimum of 437. We do not report sentence numbers, for sentence identification in legal text is a difficult, unresolved problem [4].

[4]For a sample of the documents, see http://www.legislation.gov.uk/asp/2016/3, http://www.legislation.gov.uk/asp/2016/14/part/1/chapter/1 or http://www.legislation.gov.uk/ssi/2010/407/made

The workflow was managed on Trello. The (annotated) documents were stored in shared Google Docs directories, which corresponded to the annotation steps. Github served as a code and XML repository. The XML annotated files were transfered to a web site on which they can be queried by XQuery and re-visualised using XSLT.

Some points of disagreement between annotators lead us to revise and clarify the annotation guidelines. Many questions focused on the scope of the annotations and more explicit guidelines have been given, *e.g.* in case of lists and complex sentences. The interpretations of modal verbs, like "may" or "must", also raised questions as they cannot be automatically matched to one type of prescriptive statement. Examples have been added to draw annotators' attention on these issues. In some cases, legal annotators lacked the basics of logical reasoning and needed additional explanation (*e.g.* the negation of an obligation is a permission). The annotation of reparations and exceptions appeared to be particularly difficult, probably because of the diversity of possible formulations: the guidelines have been enriched with examples and interpretation tips.

5. Results

In this section, we discuss the project outputs, which are:

- A very simple annotation language designed for legal annotators and for an automatic transformation into LegalRuleML compliant annotations.
- An annotation manual which provides 1) guidelines for the homogeneous application of legal semantic annotations and 2) instructions on the workflow.
- An annotated corpus and its corresponding LegalRuleML encoding. Presently, 558 statements are annotated.
- A dedicated web application[5], for retrieving the annotated statements based on their types as well as on the keywords or text patterns they contain.

In Figure 1, we have a snippet of source text annotated as a prohibition. Opening and closing brackets indicate the beginning and ending of the annotated text span. A number is introduced to facilitate relating expressions. In Figure 2, we provide the corresponding expression in LegalRuleML. Note that the XML structure requires auxiliary information not found in the source text with annotation, including `PrescriptiveStatement`, a (bodiless) `Rule` with conclusion `then`, a deontic element `Prohibition`, all wrapping the full text as a `Paraphrase`. Note that within `Paraphrase`, we have copied the source text. Thus, our approach maintains the source text for further analysis *in situ*, while wrapping it in valid LegalRuleML. Finally, Figure 3 presents the statement amongst the query results for both "offence" and "tobacco products" contained within a `PrescriptiveStatement` that is a `Prohibition`.

Most of the questions listed in Section 3 can be answered using the search tool:

1. All the definitions of offences involve the word "offence". Searching this word yields 70 statements of different kinds. To focus on definitions, we require also that the statement be a `Prohibition`, which reduces to 26 answers (recall 1, precision .84). Associated defenses are obtained by searching `Permission` elements which contain any of "defence" or "offence" (recall 1, precision .60). In

[5] http://tal.lipn.univ-paris13.fr/LexEx

```
<!-- Prescriptive Statement: 1 -->
<lrml:PrescriptiveStatement key="ps1">
 <ruleml:Rule>
  <ruleml:then>
   <lrml:Prohibition>
    <lrml:Paraphrase> (1) A person who in the course of business displays or causes to be
     displayed tobacco products or smoking related products in a place where tobacco products are
     offered for sale commits an offence. </lrml:Paraphrase>
   </lrml:Prohibition>
  </ruleml:then>
 </ruleml:Rule>
</lrml:PrescriptiveStatement>
```

Figure 2. LegalRuleML Representation

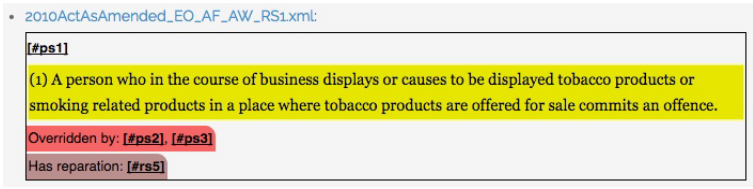


Figure 3. Query Result for "offence" and "tobacco products" in Prohibition

both cases, all the erroneously recovered statements do not specify the offence but the procedure which applies in case of offence. Last, defenses and reparations are linked to their corresponding offense *via* relations which appear on Fig 3.

2. Enumerating prohibitions which apply to tobacco products is more difficult because of alterantive lexicalisations. A search for "tobacco product[s]" in `Prohibition` elements gets 6 statements. But for "It is an offence for an adult to smoke in a private motor vehicle when there is a child in the vehicle"? terminological knowledge would help. When interpreting "A person who fails to comply with a requirement made under subsection (1) or (2) commits an offence", one needs to refer to subsections (1) and (2).
3. Obligations placed on shop owners are, for similar reasons, difficult to select. "Shop" appears only once in the texts and "owner" never, "business" being the more usual term, but also "management", "control", and "responsible person".
4. Permissions given to Scottish Ministers are easier to focus on because the title is always literally used. Querying "Scottish Ministers" in `Permission` elements yields 21 statements (precision .952, recall .875). On one side, 1 permission is given to "a person" ; on the other side, 3 additional permissions are incidentally mentioned in `Obligation` or `Constitutive` statements.
5. As can be seen in Fig 3, related overriding or reparation provisions are mentioned and accessible through a direct link in the display.

6. Discussion

Ours is not the first work to attempt the semantic annotation of legal rules, *e.g.* [6]. However, it is the first to tie the annotation effort directly to some well-developed, standardised markup language such as LegalRuleML. In our view, and following [3], the development of a high quality annotator manual which leads a team of annotators to a high level of inter-annotator agreement is an essential task in its own right. Setting up an ef-

ficient and simple workflow of annotation is also important if one wants annotators to concentrate on interpretative issues.

Returning to Figures 1-2, our methodology highlights an important issue in formalising source text: annotation requires analyzing the expression. As is apparent, we have taken the "naive" approach of annotating a whole sentence according to key words; that is, (1) is marked as a prohibition given *commits an offence*. Yet, obviously, this is misleading since the contents of the whole annotated text, including *commits an offence*, is not what is prohibited. Rather, what is prohibited is the action *displays or causes to be displayed tobacco products or smoking related products in a place where tobacco products are offered for sale* committed by *a person in the course of business*. What the search tool ought to return is just those prohibitions with respect to their content. Providing such analysis requires some care so as not to distort the meaning of the source expression. Yet, it is such fine-grained analyses that LegalRuleML requires. Our simplified, incremental approach to annotation is but one step towards this more refined result, whilst highlighting problems to address as well as yielding useful results along the way.

Finally, some of the missing results are matters beyond LegalRuleML, *e.g.* lexical semantic relationships amongst terminology. There are interesting interpretive issues concerning linguistic expressions of the annotations, complex expressions, ellipsis, reference, and others. Nonetheless, an advantage of our effort is to draw out a detailed, extensive range of such matters. Thus, there remains significant work ahead.

Now that the annotation guidelines and process have been tested and revised thanks to the adjudication work, a larger annotation experiment can be launched. The quality of the resulting annotated corpora (measured as the inter-annotator agreement) is a key feature, as our ultimate goal is to use it as training data for automating (part of) the annotation process.

References

[1] Tara Athan, Guido Governatori, Monica Palmirani, Adrian Paschke, and Adam Z. Wyner. Legalruleml: Design principles and foundations. In Wolfgang Faber and Adrian Paschke, editors, *Reasoning Web. Web Logic Rules - 11th Int. Summer School, Berlin, Germany, 2015, Tutorial Lectures*, pages 151–188. Springer, 2015.

[2] Pompeu Casanovas, Monica Palmirani, Silvio Peroni, Tom M. van Engers, and Fabio Vitali. Semantic web for the legal domain: The next step. *Semantic Web*, 7(3):213–227, 2016.

[3] Karën Fort. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Wiley-ISTE, July 2016.

[4] Jaromir Savelka and Kevin Ashley. Extracting case law sentences for argumentation about the meaning of statutory terms. In *Proc. of the 3rd Workshop on Argument Mining (ArgMining@ACL 2016)*, Berlin, Germany, 2016.

[5] Adam Wyner, Adeline Nazarenko, and Francois Lévy. Towards a high-level controlled language for legal sources on the semantic web. In Brian Davis, J. Gordon Pace, and Adam Wyner, editors, *Proc. of the 5th Int. Workshop on Controlled Natural Language (CNL2016)*, pages 92–101, Aberdeen, UK, July 2016. Springer.

[6] Adam Wyner and Wim Peters. On rule extraction from regulations. In Katie Atkinson, editor, *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, pages 113–122. IOS Press, 2011.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-107

Balancing with Thresholds

Michał ARASZKIEWICZ [a,1], Tomasz ZUREK [b]
[a] *Departament of Legal Theory, Faculty of Law and Administration, Jagiellonian University, Ul. Gołębia 24, 31-007 Kraków, Poland*
[b] *Institute of Computer Science, Maria Curie Sklodowska University, Ul. Akademicka 9, 20-033 Lublin, Poland*

Abstract The paper presents a general formal framework representing the role of balancing of values in interpretation of statutory rules. The model developed here is an extension of the model of teleological interpretation, where a given interpretive outcome is justified if it satisfies a given goal (or a set of goals). Herein, a richer argumentative structure is discussed: an interpretive proposition concerning the interpretation of a statutory condition is justified if it is in accordance with the proper balance of applicable legally relevant values.

Keywords. Argumentation, Balancing, Goal-based reasoning, Statutory Interpretation, Value-based reasoning

1. Introduction

The topic of balancing of values and its role in legal reasoning has been the point of interest in AI and Law for more than two decades now. However, so far the fundamental concepts and inference patterns related to balancing have not been accounted for in a formal framework in the context of statutory interpretation. This paper aims to fill this gap. The paper does not deal with the structure of the balancing itself, but it argues for a basic conceptual scheme that creates the background for any instance of balancing-based interpretation. The results of our work may be useful for the development of rule-based systems involving the notion of interpretation.

In legal literature the topic of balancing has been initially associated with the notion of legal principles [7], [1], [2], [10], [3]. In the domain being the scope of this paper – that is, statutory interpretation – the issue in question is whether a given rule should be interpreted in certain manner and, as a consequence, applied to the given state of affairs. Therefore, the objects being valuated with respect to relevant values are states of affairs with attached consequences following from the rule in question, and, for comparison, the same states of affairs without such consequences (similarly to [12] and [8]). As far as the criteria of acceptance are concerned, two aspects have to be distinguished. First, no legally relevant value should be realized below its core threshold [10]. Second, it is an open question whether we are obligated to adopt the interpretation which yields the optimal level of balancing of values, or is it acceptable to adopt any outcome which satisfies

[1]Corresponding Author: Departament of Legal Theory, Faculty of Law and Administration, Jagiellonian University, Ul. Gołębia 24, 31-007 Kraków, Poland; E-mail: michal.araszkiewicz@uj.edu.pl The writing of this paper was supported by the Polish National Science Centre (research project No DEC-2015/17/B/HS5/00457)

a certain valuation threshold. Our thesis is that this threshold of minimal acceptability is typically fixed by means of interpretive propositions based on balancing. Our paper is based on the idea which can be seen as the development and discussion of the teleological interpretation concept from [12], but also as an implementation of the concept of goal and relationship between goal and value from [11].

2. The model

We will begin with a summarized discussion of the basic concepts of the model of teleological reasoning from [11], further referred to as the GVR model:
Let $S = \{s_x, s_y, s_z, ...\}$ be a finite, non-empty set of propositions. Each proposition represents one state of affairs. We have to separate the two meanings of the word value: a value may be understood as a concept or as a process: (1) Value as an abstract concept which allows for the estimation of a particular action or a state of affairs and influences one's behaviour. V is a set of values: $V = \{v_1, v_2, \dots v_n\}$ (2) Valuation as a process of estimation of the level of extent to which a particular states of affairs s promotes a value v_i. By $v_i(s)$ we denote the extent to which s promotes a value v_i. By $V(S)$ we denote the set of all valuations of all states of affairs. By $V^i(S)$ we denote the set of all possible extents to which a value v_i from set V may be promoted by any possible state of affairs $s \in S$. A partial order $O_i = (\geqslant; V^i(S))$ represents the relation between extents to which values are promoted.In real-life reasoning people do not rely only on a comparison of the levels of promotion of one value; usually, they compare the levels of promotion of various values. Theoretically speaking, they are incompatible, but practically, people compare not only the levels of promotion of various values, but also the levels of promotion of various sets of values. By $V^Z \subset V$ we denote a subset (named Z) of a set of values V which consists of values: $v_i, v_j, ... \in V^Z$. By $V^{s_i} \subset V$ we will denote a set of values promoted by a state of affairs s_i.
By $V^Z(s_n)$ we denote a set of estimations of the levels of promotion of values constituting set V^Z by a state of affairs $s_n \in X$. If $V^Z = \{v_z, v_t\}$, then $V^Z(s_n) = \{v_z(s_n), v_t(s_n)\}$. A partial order $OR = (\rhd; 2^{V(S)})$ represents a preference relation between various sets of values and various states of affairs: $V^Z(s_n) \rhd V^Y(s_m)$ means that the extent to which values from set V^Z are promoted by a state of affairs s_n is preferred to the extent to which values from set V^Y are promoted by a state of affairs s_m.
The discussion of relationships between orders OR and O as well as the mechanism of deriving order OR is presented in the [11].

Definition 1 (Legal rule) *Let $R = \{r_v, r_z, ...\}$ be a set of legal rules. Each rule is a pair $\langle s_x, c_x \rangle$, where s_x is the condition of the rule and c_x is the conclusion of the rule.*
If a state of affairs s_a fulfills the conditions of the rule, then the conclusion leads to the change of the state of affairs into s_{x+c} (where $s_x, s_{x+c} \in S$).

Basing on [10] we assume that the grounds for evaluation of each interpretation in our model will be goals in the form of minimal extents to which a given set of values should be promoted. Thus established concept of goal remains complacent with the idea of abstract goal from [11]:

Definition 2 (Goals) *Goals are represented by the minimal acceptable extents to which a particular state od affairs promotes a given set of values:*

Let $GA = \{ga_1, ga_2, ...\}$ be a set of goals. By $v_n min(ga)$ we denote the minimal extent to which the promotion of a value v_n satisfies a goal ga. By $v_n(s_1) \geqslant v_n min(ga)$ we denote that a goal ga is satisfied by a state of affairs s_1 with respect to a value v_n. By $v_n \in ga$ we denote that the minimal extent of a given value v_n is declared in a goal ga (note that $\in$ is different that $\in$). The abovementioned definition of goals represents the idea of protection of the core of values; the defined goals correspond to core thresholds as discussed in [10]. Other types of goals are also relevant in law, but the minimal thresholds are particularly important, defining the minimal acceptability of statutory interpretation statements.
Although like Sartor we assume that the foundation for setting goals are values whose promotion is recommended by principles (goal norms).

Definition 3 (Interpretation) *The binary operator $\bullet$ represents interpretation of the principle's conditions (the operator $\bullet$ was extensively discussed in [4] and [5]). By $s_t \bullet s_x$ (where $s_t, s_x \in S$) we mark that a state of affairs s_t fulfills conditions s_x.*

It should be pointed out that in our model we introduced a differentiation between the current state of affairs and the state of affairs expressed in the rule's premises. It results from the fact that in practice the description of the actual state of affairs very rarely literally matches the premises of the rule; most frequently it is somehow interpreted, often by the so-called intermediate legal concepts.

Definition 4 (Interpretive Statements) *All complex expressions of the elements of set S and constructed by means of the relation word $\bullet$ will be referred to as Interpretive Statements.*

Interpretive Statements play a role of intermediaries between the factual description of a given state of affairs and the states of affairs expressed in the conditions of legal rules. The crucial question is this context is whether the conditions of a rule should be interpreted in such a way to encompass the current fact situation, or to the contrary. Interpretive canons [9] serve as arguments for justification of this or another Interpretive Statements concerning the conditions of the rule in question. However, these canons may also be looked at as heuristics: simplified rules approximating the actually justified Interpretive Statements. If we agree that law is a system designed for the sake of realization of important social values, then we may assume that the set of "actually justified" Interpretive Statements follow from the balancing of those socially relevant values. Note that we do not claim the existence of a unique "right" interpretation of any legal rule [7].
Let us now consider application of a given rule $r_a\langle s_a, c_a\rangle$ to the state of affairs s_m. In order to justify this application, we have to be able to show that $s_m \bullet s_a$ (for the sake of simplicity we do not consider the problem of analogous application of rules here). Typically, justifying this inference step will involve at least one layer of intermediary concepts.
Let us define the set $IS\ (r_a, s_m)$ as the set of Interpretive Statements concerning apppplication of r_a to s_m.

Definition 5 (Positive and Negative Interpretive Arguments) *An Interpretive Statement $\in IS(r_a, s_m)$ is a Positive Interpretive Statement ($PINS$) if and only if it justifies application of r_a to s_m. An Interpretive Statement $\in IS(r_a, s_m)$ is a Negative Inter-*

pretive Statement Statement ($NENS$) if and only if it justifies non-application of r_a to s_m.

Definition 6 (Goal-admissible Interpretive Statement) *An interpretive statement $\in IS(r_a, s_m)$ is goal-admissible with respect to goal ga warranting the realization of value v_n if and only if:*

a. $v_n(s_m) \geqslant v_n min(ga)$ if the interpretive statement is a $NENS$ (goal- admissible $NENS$)
b. $v_n(s_{m+ca}) \geqslant v_n min(ga)$ if the interpretive statement is a $PINS$ (goal - admissible $PINS$)

If an interpretive statement is not goal-admissible, then it is goal-inadmissible.

It is worthwhile to delimit the set of states of affairs in which each Interpretive Statement, concerning application of legal rules R, is goal-admissible with respect to the value v_n and the goal ga, setting its minimal required realization. We will refer to this set as the Model of Interpretive Statements with regard to value v_n and goal g_a.

Definition 7 (Model of Interpretive Statements – single value) . *Let IS be the given set of Interpretive Statements, v_n the value in question and ga – the goal protecting the minimal realization of the value. $Mod(IS)_{v_n,ga}$ is the set of all states of affairs in which all given Interpretive Statements are goal-admissible.*

Intuitively, the model of interpretive statements is the set of all states of affairs where the degree of realization of a value is always greater than the minimal threshold, taking into account the rules applicable to these states of affairs and interpretation of these rules. A given state of affairs may not belong to the model of interpretive statements if this state of affairs is untypical or novel (hard cases), or if the legislation is poorly drafted, enabling goal-inadmissible interpretations.
Conversely, we may define the set of all goal-admissible interpretive statements for a given set of states of affairs, taking into account value v_n and goal ga. We will refer to this set as the Interpretive Theory ($INTh$) of a set of states of affairs.

Definition 8 (Interpretive Theory of States of Affairs – single value) *Let S be the given set of States of Affairs, v_n the value in question and ga – the goal protecting the minimal realization of the value.*
$INTh(S)_{v_n,ga}$ is the set of all Interpretive Statements that are goal-admissible in any of the states of affairs in S.

The operator $INTh(S)_{v_n,ga}$ separates the goal-admissible from the goal-inadmissible interpretive statements, for a given set of states of affairs, taking into account the realization of v_n with regard to ga.
Let us now combine the two operators to obtain the notion of Value-based Consequence of the given set of Interpretive Statements.

Definition 9 (Value-Based Consequence) *Value-based consequence of the set of Interpretive Statements $VCn(IS)_{v_n,ga}$ is defined as $INTh_{v_n,ga}(Mod(IS)_{v_n,ga})$, that is, the set of all Interpretive Statements that are goal-admissible in all non-hard cases.*

The above definitions may be generalized to encompass sets of values and goals. Intuitively, $VCn(IS)_{v_n,ga}$ defines a relatively narrow subset of Interpretive Statements.

3. Argumentation schemes

We have already noticed ([9], [4]) that interpretive statements are justified by means of interpretive canons, usually expressed as argumentation schemes which are forms of argument which represent stereotypical patterns of human reasoning.
Below we present two interpretive canons: the first one justifies a positive interpretive statement on grounds that it fulfill the goal set by the legislator; the second one constitutes a demonstration of balancing-based interpretive conflict solution.

IAS1 The first type of argumentation scheme: every positive interpretive statement which fulfills the goal is justified. The given data are: a goal ga_k, a current state of affairs (s_m), a legal rule r_l, and a Positive Interpetive Statement: $s_m \bullet s_l$. If after the application of rule (s_{m+c}), s_m will promote all values indicated by goal ga_k to a no lesser degree than the minimum, then the interpretive statement $s_m \bullet s_l$ will be justified:

$$\frac{\begin{array}{c} ga_k \\ s_m \\ r_l = \langle s_l, c \rangle \\ s_m \bullet s_l \in PINS \\ \forall_{v_n \in ga_k} v_n(s_{m+c}) \geqslant v_n min(ga_k) \end{array}}{s_m \bullet s_l}$$

IAS2 The second argumentation scheme refers directly to the balancing of values and is an example of a conflict resolution mechanism: there are two exclusive interpretive statements, both fulfilling the set goal, but one of them is preferred because of the values it promotes: The given data are: a goal ga_k, a current state of affairs (s_m), a legal rule r_l, two interpretive statements $s_m \bullet s_l \in PINS$ and $s_m \not\bullet s_l \in NENS$. Both cases promote values indicated by ga_k to the extent no lesser than the recommended minimum. If in the context of set V^k and after the application of rule (s_{m+c}), s_m will be preferred to interpretation $s_m \not\bullet s_l$, then the interpretive statement $s_m \bullet s_l$ will be justified. The crucial point in the discussion is the list of values V^k determining on the basis of which values the balance should be made. Obviously, not all of the values ought to be taken into consideration. In previous sections we assumed that our goal (ga_k) is set on the basis of the binding legal principles; since they define the constitutional order, we believe that they should serve as the foundation of balancing. Therefore we assume that $V^k = \{v_n | v_n \in ga_k\}$.

$$\frac{\begin{array}{c} ga_k \\ s_m \\ r_l = \langle s_l, c \rangle \\ s_m \bullet s_l \in PINS \\ s_m \not\bullet s_l \in NENS \\ \forall_{v_n \in ga_k} v_n(s_{m+c}) \geqslant v_n min(ga_k) \\ \forall_{v_n \in ga_k} v_n(s_m) \geqslant v_n min(ga_k) \\ V^k(s_{m+c}) \rhd V^k(s_m) \end{array}}{s_m \bullet s_l}$$

Let us note that the argumentation schemes presented above do not have to lead to the conclusion concerning uniqueness of the best interpretive sentence (one right answer, see [9]), because set ordered by the symbol ($\rhd$) may not have the greatest element, it may have more than one maximal elements.

4. Discussion and Conclusions

In this paper we have provided a general framework which extends the concept of teleological reasoning to represent the role of balancing of values in the context of statutory interpretation. This framework is compatible with the findings of [10]. We have also defined certain specific concepts representing safe interpretive situations, where explicit balancing is not needed to justify a satisfactory interpretive outcome (the notions of model of Interpretive Statements, interpretive theory of states of affairs and finally, value-based consequence). For situations where actual balancing needs to be made explicit, we have provided two argumentation schemes.

As for the future work, we intend to: (1) explore the structure of arguments supporting conclusions encompassing the ordering operators (orders O_i and OR); this line of reserach involves investigations into case-based reasoning structures in statutory interpretation; (2) apply the framework to model the situation of justified violation of rules (as in Bench Capon [6]) and (3) integrate the model into the broader framework modeling the behaviour of agents interpreting statutes.

References

[1] R. Alexy. *A Theory of Constitutional Rights*. Oxford Press, 2003.

[2] Michal Araszkiewicz. Balancing of legal principles and constraint satisfaction. In *Legal Knowledge and Information Systems - JURIX 2010: The Twenty-Third Annual Conference on Legal Knowledge and Information Systems, Liverpool, UK, 16-17 December 2010*, pages 7–16, 2010.

[3] Michał Araszkiewicz. Argument structures in legal interpretation: Balancing and thresholds. In Thomas Bustamante and Christian Dahlman, editors, *Argument Types and Fallacies in Legal Argumentation*, pages 129–150. Springer International Publishing, 2015.

[4] Michal Araszkiewicz and Tomasz Zurek. Comprehensive framework embracing the complexity of statutory interpretation. In *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference, Braga, Portual, December 10-11, 2015*, pages 145–148, 2015.

[5] Michal Araszkiewicz and Tomasz Zurek. Interpreting agents. In *Legal Knowledge and Information Systems - JURIX 2016: The Twenty-Ninth Annual Conference*, pages 13–22, 2016.

[6] Trevor J. M. Bench-Capon. Value-based reasoning and norms. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, pages 1664–1665, 2016.

[7] R. Dworkin. *Taking Rights Seriously. New Impression with a Reply to Critics.* Duckworth, 1978.

[8] Matthias Grabmair and Kevin D. Ashley. Argumentation with value judgments an example of hypothetical reasoning. In *Proceedings of the 2010 conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference*, pages 67–76, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press.

[9] F. Macagno, D. Walton, and G. Sartor. Argumentation schemes for statutory interpretation. In M. Araszkiewicz, M. Myška, T. Šmejkalova, J. Šavelka, and Škop M., editors, *Argumentation. International Conference on Alternative Methods o Argumentation in Law.*, pages 61–76, Brno, 2012.

[10] G. Sartor. Doing justice to rights and values: teleological reasoning and proportionality. *Artificial Intelligence and Law*, 18, 2010.

[11] Tomasz Zurek. Goals, values, and reasoning. *Expert Systems with Applications*, 71:442 – 456, 2017.

[12] Tomasz Zurek and MichałAraszkiewicz. Modeling teleological interpretation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, pages 160–168, New York, NY, USA, 2013. ACM.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-113

Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links

Tommaso AGNOLONI [a], Lorenzo BACCI [a], Ginevra PERUGINELLI [a],
Marc van OPIJNEN [b], Jos van den OEVER [b],
Monica PALMIRANI [c], Luca CERVONE [c], Octavian BUJOR [c],
Arantxa ARSUAGA LECUONA [d], Alberto BOADA GARCÍA [d],
Luigi DI CARO [e], Giovanni SIRAGUSA [e]

[a] *Institute of Legal Information Theory and Techniques (ITTIG-CNR)*
[b] *Publications Office of the Netherlands (UBR|KOOP)*
[c] *CIRSFID - University of Bologna*
[d] *General Council of the Judiciary - CENDOJ*
[e] *Computer Science Department - University of Torino*

Abstract. In this paper we present the BO-ECLI Parser, an open framework for the extraction of legal references from case-law issued by judicial authorities of European member States. The problem of automatic legal links extraction from texts is tackled for multiple languages and jurisdictions by providing a common stack which is customizable through pluggable extensions in order to cover the linguistic diversity and specific peculiarities of national legal citation practices. The aim is to increase the availability in the public domain of machine readable references metadata for case-law by sharing common services, a guided methodology and efficient solutions to recurrent problems in legal references extraction, that reduce the effort needed by national data providers to develop their own extraction solution.

Keywords. natural language processing, legal references, case law databases, linked open data

1. Introduction

Among the goals of the European Case Law Identifier (ECLI) established in 2010[1] is the publication of national case-law by courts of European member States via the ECLI Search Engine on the European e-Justice Portal. Besides being uniformly identified, decisions should be equipped with a minimal set of structured metadata describing their main features. Among the (optional) metadata prescribed by the ECLI Metadata Scheme, *references* metadata describe relations of the current document with other legal (legislative or judicial) documents, formally expressed using uniform identifiers (the aforementioned ECLI for case-law, ELI for legislation, national identifiers, CELEX identifiers for European legal documents). These relational metadata are at the same time among the

[1] Council conclusions inviting the introduction of the European Case Law Identifier (ECLI) and a minimum set of uniform metadata for case law (CELEX:52011XG0429(01)).

most useful case-law metadata - in that they allow the enhancement of legal information retrieval with relational search - and among the most difficult to have valued, especially for legacy data and for less resourced languages and jurisdictions. While manual reference tagging is an extremely costly procedure - not viable in the public domain and especially to cope with the growing amount of data published in national case law databases - automatic legal reference extraction has been successfully applied in several national contexts [1], [2], [3] despite the complexity of coping with a diversity of styles, variants and exceptions to existing drafting rules and citation guidelines.

Based on an analysis of approaches and existing solutions to the "Linking data" problem [4] and on the results of a survey on citation practices within EU and national Member States' courts [5], the BO-ECLI Parser presented in this work and developed within the EU funded project "Building on ECLI"[2], tackles the problem from a EU-wide multi-lingual / multi-jurisdictional perspective. With a strong commitment to openness (open source software, open data, open formats) the aim is to reduce the effort for national data providers willing to develop their own legal reference extraction solution by sharing a proven methodology and efficient solutions to recurrent problems in legal references extraction.

The BO-ECLI Parser is structured as an architecture of interoperable services (Fig. 1). The core of the extraction process is taken care of by the Parser Engine (Sect. 2). The REST API exposes the results of the reference extraction process as structured interoperable XML and JSON open formats (Sect. 3). Data Services provide access to authoritative repositories of legal references allowing to complement the informations extracted by the Engine (Sect. 4). An extensible User-Interface is also provided for direct user interaction and as a proof of concept of the integration of the different services (Sect. 5).

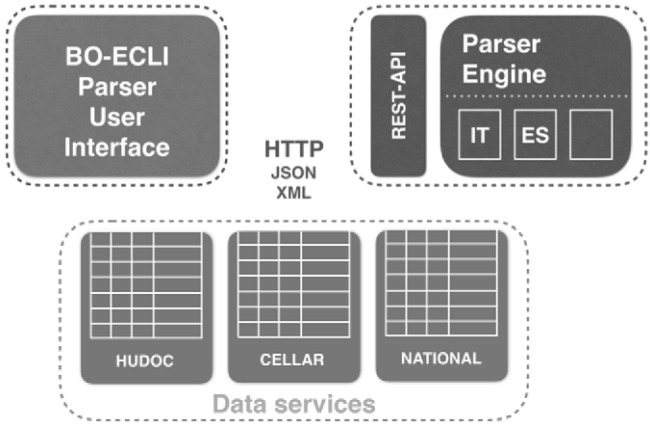


Figure 1. The overall architecture of the BO-ECLI Parser.

[2]http://bo-ecli.eu

2. Parser Engine

The BO-ECLI Parser Engine [6] is an extensible framework for the extraction of legal links from case-law texts. It is written in Java and distributed as open source software[3]. It targets citations to both case-law and legislation, expressed as lists of textual features (authority, type of document, document number, date, etc.) or as common names (i.e. aliases). Multiple citations, intended either as citations to more than one partition of a single document or as citations to more than one document issued by a single authority, are also covered and distinct legal references are generated in correspondence to each partition and each document. A distinguishable characteristic of the software consists in the capability to be extended in order to support the extraction process from texts written in different languages or issued within different jurisdictions.
In order to realize such design, two practical steps are required:

- dividing the process of legal link extraction into a generic and customizable sequence of atomic services, following a pipeline pattern;
- defining an annotation system able to convey the work done by each service along the pipeline.

2.1. A pipeline of services

One way to synthesize a generic process of legal link extraction from texts is, first, to divide it into three consecutive phases:

1. the entity identification phase, where the fragments of text that can potentially represent a feature of a citation are identified and normalized;
2. the reference recognition phase, where patterns of identified features are read in order to decide whether they form a legal reference or not;
3. the identifier generation phase, where the recognized legal references are analyzed so that standard identifiers, and possibly URLs, can be assigned to them.

Secondly, within every single phase, a number of different services can be placed, each specialized in absolving one task. For example, within the entity identification phase, there could be a service specialized in the identification of case numbers.

2.2. Annotation system

The BO-ECLI Parser Engine framework defines an internal annotation system to allow every service implementation, especially the ones belonging to entity identification and reference recognition, to save the specific results of their execution directly in the text.
Annotations are used to assign a category (hence, a meaning) to a fragment of text, while, through normalization, annotated fragments of text can acquire a language independent value. For example, the Italian fragment of text *"sent. della Corte Costituzionale"*, meaning a judgment issued by the Italian Constitutional Court, at a certain point along the pipeline, is annotated as follows:

```
[BOECLI:CASELAW_TYPE:JUDGMENT]sent.[/BOECLI] della
[BOECLI:CASELAW_AUTHORITY:IT_COST]Corte Costituzionale[/BOECLI]
```

[3] http://gitlab.com/BO-ECLI/Engine

Thanks to the annotation system, the work of each service is conveyed and shared along the pipeline in a language independent way.

2.3. Service implementation

The implementation of an annotation service belonging to either the entity identification or the reference recognition phase simply consists in a piece of software that analyzes an input text, possibly already enriched with annotations, and produces an equivalent output text, possibly with altered annotations. The default implementations of the annotation services provided by the framework make use of JFlex[4], a well-known lexical scanner generator for Java.

A number of implementations for services that belong to each phase of the legal link extraction process are provided by the framework by default. Typically, a default implementation is supplied when the task that the service is in charge of can be considered language independent, pertains to the European jurisdiction or is common in the European context.

Parties identification: The identification of the names of the parties in a citation should be generally considered as a language dependent task. Nonetheless, the framework provides a default service implementation for the identification of applicants and defendants relying on heuristics based on positioning, upper and lower casing, the *versus* entity and the geographic identification of a country member of the Council of Europe (as a defendant in European Court for Human Rights citations).

Reference recognition: After the entity identification phase, the textual features that can potentially be part of a legal reference are annotated and normalized, hence they can be treated as language independent entities. Although citation practices change from one jurisdiction to another, the framework provides a number of default service implementations for reference recognition that are able to cover the most typical citation patterns and, also, to support multiple citations.

ECLI generation for European Courts: In those cases where a standard identifier can be simply generated as a composition of the features extracted from the textual citation, the framework provides a default service implementation to automatically assign an identifier to a legal reference. This is the case for the generation of ECLI for legal references that have the European Court of Human Rights as the issuing authority, when the type of document, the case number and the date are known.

CELEX generation for European legislation: Another service implementation supplied by the framework for the automatic composition of a standard identifier is used for legislation references to European directives and regulations. For these types of document, when the referred document number and year are known, a CELEX identifier as well as its ELI identifier can be assigned to the legal reference.

3. REST-API and structured reference exchange format

A REST API is wrapped around the Java API of the Engine in order to allow its exposition as a service on the Web via the HTTP protocol and to guarantee interoperability with additional components possibly written in different languages. The Engine REST-API

[4] http://jflex.de

exposes the results of the reference extraction process performed by the Engine as structured XML and JSON open format for their consumption by additional services and for the possible further enrichment and validation of the results of the automatic extraction.

The API response provides, for each text fragment where a citation has been detected, a structured representation of the corresponding reference, listing its attributes along with their normalized values. In case of multiple citations (in the sense described in Sect. 2) a collection of references is returned each associated with the corresponding text fragment.

4. Open Data Services

For those cases where the identifier cannot be computed by the composition of the reference features used in the textual citation, it is mandatory to look-up such standard identifiers (preferably European standard identifiers: ECLI for case-law and ELI and for legislation) by querying reference catalogs. In the BO-ECLI Parser design this is accomplished by reusing existing reference repositories possibly exposed as Open Data on the web accessible via HTTP APIs.

Due to their importance to all national jurisdictions, two data services have been implemented to get standard identifiers of references to case-law issued by the Court of Justice of the European Union (CJEU) and by the European Court of Human Rights (ECHR) for which ECLIs cannot be straightforwardly computed based on the features and numbering typically used in textual citations.
Additionally, national reference repositories can be reused and integrated in order to accomplish national identifiers look-up. Standardizing the access to such metadata repositories through a common layer is among the long term goals of the BO-ECLI Parser framework [7].

5. User-Interface

Though the parser is primarily intended to be used through its API for integration in different systems, an extensible open-source User Interface developed using modern Web technologies (Node.js) is also provided for direct user interaction. The UI interacts with the different Web services through HTTP and provides a proof of concept of their integration. Functionalities are provided to set the input text and parameters and inspect the results extracted by the BO-ECLI Parser in different views: annotated HTML text, tabular view, structured exchange format (JSON) view for developers, *references* metadata according to the official ECLI Metadata Scheme. The UI project is extendable to the needs of the national judiciary for testing or production, e.g. for manual check and validation of the results of the automatic extraction before the deposit in a case law management and publication system. A deployed demo version of the UI is accessible as part of the website of the BO-ECLI project[5].

[5] http://parser.bo-ecli.eu

Conclusions

We presented the BO-ECLI Parser, an open source framework for the automatic extraction of case-law and legislation references from case-law texts issued in the European context. Its architecture is based on the interaction of different interoperable and extensible components. In particular, the Parser Engine provides a framework where national extensions can be developed and plugged in order to add support for the extraction process from texts written in different languages or issued within different jurisdictions. By defining and providing a complete stack for legal links extraction, the implementation of a national extension is guided and straightforward and the effort needed for the development of a fully functional national extractor is considerably reduced. Along with the framework project, a Template project has been developed in order to facilitate and encourage the adoption of the software for the extraction of legal links in new languages and jurisdictions. Two concrete national extensions have been developed so far by different teams to support the extraction from Italian and Spanish case-law texts, proving both the feasibility and the straightforwardness of the whole approach. The BO-ECLI Parser software projects, their code and documentation are hosted on the GitLab software development platform[6].

Acknowledgement

This publication has been produced with the financial support of the Justice Programme of the European Union. The contents of this publication are the sole responsibility of the authors and can in no way be taken to reflect the views of the European Commission.

References

[1] Lorenzo Bacci, Enrico Francesconi and Maria Teresa Sagri. A Proposal for Introducing the ECLI Standard in the Italian Judicial Documentary System. In *Proceedings of the 2013 Conference on Legal Knowledge and Information Systems: JURIX 2013: The Twenty-sixth Annual Conference* pp. 49-58 IOS Press Amsterdam (NL), 2013.

[2] Marc van Opijnen, Nico Verwer and Jan Meijer. Beyond the Experiment: The Extendable Legal Link Extractor. In *Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, June 8-12 2015 held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL) San Diego, CA, USA. Available at SSRN: https://ssrn.com/abstract=2626521.

[3] A. Mowbray, P. Chung and G. Greenleaf. A free access, automated law citator with international scope: the LawCite project. *European Journal of Law and Technology* **7** (3), 2016. Available at: http://ejlt.org/article/view/496/691 .

[4] Tommaso Agnoloni and Lorenzo Bacci. *BO-ECLI project deliverable D2.1 Linking Data - analysis and existing solutions*, 2016. Available at: http://bo-ecli.eu/uploads/deliverables/DeliverableWS2-D1.pdf .

[5] Marc van Opijnen, Ginevra Peruginelli, Eleni Kefali and Monica Palmirani. *On-line Publication of Court Decisions in the EU - Report of the Policy Group of the Project 'Building on the European Case Law Identifier'*, 2017. Available at: http://bo-ecli.eu/uploads/deliverables/Deliverable%20WS0-D1.pdf .

[6] Tommaso Agnoloni, Lorenzo Bacci and Marc van Opijnen. BO-ECLI Parser Engine: the Extensible European Solution for the Automatic Extraction of Legal Links. In *Proceedings of the 2nd Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, June 16 2017 held in conjunction with the 2017 International Conference on Artificial Intelligence and Law (ICAIL) London, UK.

[7] Marc van Opijnen, Monica Palmirani, Fabio Vitali, Jos van den Oever and Tommaso Agnoloni. Towards ECLI 2.0. In *CeDEM17 Proceedings of the International Conference for E-Democracy and Open Government*, May 17-19 2017. Danube University Krems, Austria.

[6] https://gitlab.com/BO-ECLI

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-119

Scoring Judicial Syllabi in Portuguese

Jean-Rémi BOURGUET [a,1] and Melissa ZORZANELLI COSTA [b,2]
[a] *Núcleo de Estudos em Modelagem Conceitual e Ontologias*
Federal University of Espírito Santo (UFES) – Brazil
[b] *Tribunal Regional Federal da 2ª Região*
Justiça Federal - Seção Judiciária do Espírito Santo (JFES) – Brazil

Abstract. Law professionals generally need to investigate a large number of items to make their decisions. However, the frameworks they use are often limited to a simple full-text search. In this paper, we propose to score the results of such searches investigating ontological and non-ontological solutions. We examine their applicabilities in a real use case dealing with jurisprudences of regional federal courts in Brazil.

Keywords. Jurisprudences, Full-text search, NLP, Portuguese, Similarities

1. Introduction

Nowadays, more and more Application Programming Interfaces (APIs) supporting a Natural Language Processing (NLP) are released and their mutual usages in common infrastructures can considerably enhance the results of full-text searches. As they are regularly confronted with large numbers of judged cases stored in relational data management systems, professionals of Brazilian courts need innovations to refine these results. Indeed, if the usual way to output a full-text search is an ordered list of items, few approaches are thought to display the results in different ways. Actually, the jurisprudences that are judicial decisions taken by a specific court in Brazil (e.g. regional federal tribunal), are stored in semi-structured formats in which a large part of the relevant knowledge are present in syllabi, i.e. textual explanations in Portuguese. Our proposal is then to score the results of a full-text search among judicial syllabi supported either by ontological or non-ontological solutions. On the one hand, we took up the challenge to perform some automatic translations of the syllabi into English (with Google Cloud Translation) before computing similarities from the Princeton WordNet [1]. On the other hand, we opted to proceed word embeddings of the Brazilian penal code using Word2vec for Lucene [2] before computing similarities between lemmas by cosine measures. The remainder of this paper is organized as follows: Section 2 describes the ontological and non-ontological solu-

[1]Corresponding Author: Jean-Rémi Bourguet (jean-remi.bourguet@ufes.br) is supported by the Brazilian Research Funding Agency FAPES (grant 71047522).
[2]Melissa Zorzanelli Costa (mzcosta@jfes.jus.br) would like to acknowledge the assistance of the Tribunal Regional of 2nd Region of Brazil.

tions to score such results and evaluates the applicability of our approach through a real use case and an interface, Section 3 mentions the most similar approaches and Section 4 concludes and opens some research perspectives.

2. Scoring the results

We introduce in Definition 1 a possible global score between two texts as the maximal similarity score between their components. We denote $\overline{w_k}$ the lemmatization of a word w_k. We arbitrarily avoided taking into account the score with words present in a stop words set denoted θ.

Definition 1. *Let two texts p, q:*

$$SIM\,(p,q) = \max_{p_i \in p \setminus \theta, q_j \in q \setminus \theta} sim(\overline{p_i}, \overline{q_j})$$

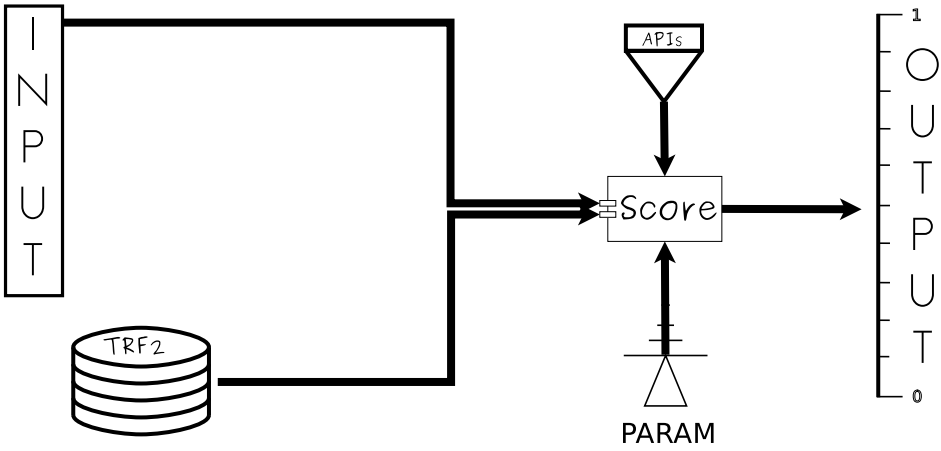


Figure 1. System description

Looking at the system description of Figure 1, several modules supports the computation of the similarity scores: `INPUT` takes a text in Portuguese, `TRF2` is a data set of jurisprudences, `APIs` is a set of NLP-based APIs supporting the computations of the similarity scores, `PARAM` manages the choice of the local measures, the stop words list, etc., `SCORE` orchestrates the computation of the similarity scores and `OUTPUT` is devoted to display them.We transformed the XML file provided by a regional Brazilian federal tribunal in an RDF file using an XSLT transformation. After that, we stored this knowledge base in a triple store supported by the Virtuoso's infrastructure [3]. We used the Jena API to query Virtuoso possibly from a Java interface developed with the native Swing API.

Performing similarities in a language other than English can be a challenge. Then, we will present both ontological (in Section 2.1) and non-ontological (in Section 2.2) approaches using semantic similarities or word embeddings to score the results of a full-text search.

2.1. Using semantic similarities

Semantic similarities generally use distances between semantic units (called synsets) in a ontologically founded lexicon. According to Sartor et al. [4], a synset can be defined as a set of one or more uninflected word forms (or lemmas) called word-senses belonging to the same part of speech (denoted pos) e.g. noun, verb or adjective. Each synset is encoded with its *lemma*, *pos* and a number *nb* discriminating it among some other possible senses (lemma#pos#nb). Finally, the relation of hyponymy (denoted $<^h$) is a binary relation that partially orders the synsets. Currently, the Global WordNet Association indexes a large set of open-source adaptations of lexicons in approximatively 40 languages. Concerning the Portuguese language, two repositories are referenced: ONTO.PT [5] a Portuguese repository automatically built from heterogeneous textual sources, and OpenWN-PT [6] a Brazilian Portuguese repository built using Wikipedia and alignments with the Princeton WordNet (conserving its abstract relations), manual revisions and gloss translations. Unfortunately, no API exists in the market to equip ONTO.PT or OpenWN-PT with a support for the computation of semantic similarities. Therefore, we opted for an automatic translation of each syllabus (and the input) in English using Google Cloud Translation. After that, we were able to perform semantic similarities with WordNet:Similarity (developed by Pedersen et al. in [7] and redesigned in Java by Shima [8]). One of the most common measures is the path-length relatedness founded on a node-counting scheme concerning the smallest specified role counting between two synsets (see [7]). The path-length based relatedness score (`plr`) is equal to the inverse of the shortest path length between two synsets. Nevertheless, other historical similarity measures considering the maximum depth in taxonomy (`lch`), the depth of the least common subsumer (`wup`) or the supported information content (`jcn` and `lin`) can also be computed with WordNet:Similarity.

2.2. Using word embeddings

Word embeddings is a well-known approach based on deep learning (using a neural natural language model) recently retailored by Mykolov et al. [9]. These similarities arise from a large set of word vectors produced after a learning step performed on a textual corpus ideally in a particular context of interest. This corpus may receive a pretreatment (e.g. tokenization, lemmatization, stop word filtering) in order to decrease the noise of nonsense textual information. Thus, the learning step encodes the general context of words in dense vectors. The similarity between two lemmas is obtained through the cosine of their vectors. We opted to use the Brazilian penal code[3] to perform the learning step. We used an API called Word2vec for Lucene [2] and proceeded to a lemmatization of the text (using Lematizador [10]). After that, we filtered the corpus with stop words, stop signs and numbers giving a train file of 11593 lemmas. We performed word embeddings using a size 200 for the vectors, a window (max skip length between words) of 5, discarding words that appear less than 5 times. We finally obtained a vocabulary size of 589 lemmas.

[3] http://www.planalto.gov.br/ccivil_03/decreto-lei/Del2848compilado.htm

2.3. Evaluation

We performed an evaluation of our framework looking for the word *banco* (i.e. bank in English) to quickly browse the large repository of jurisprudences by outputting an affordable set. Afterwards, we scored the answer by using semantic similarities (`PLR`) or by using word embeddings (`EMB`) with the word *dano* (i.e. damage in English). In the Table 1, we related the 4 most similar jurisprudences with the 3 highest local similarities scores obtained by semantic similarities and word embeddings. The first column describes a part of the Syllabi with the most similar words written in bold while the other columns show for each approach the most similar words, their scores and their translations between parenthesis.

Syllabus	`PLR`	`EMB`
DIREITO CIVIL. ATIVIDADE BANCÁRIA [...] Tal responsabilidade somente fica descaracterizada na **ocorrência** de uma das hipóteses do § 3º do referido art. 14, o que não **ocorreu** na espécie. 2 - O princípio da reparabilidade do **dano** moral foi expressamente reconhecido [...] 5 - Em face da responsabilidade **civil** contratual, aplicável a inversão do ônus da prova prevista no artigo 6º [...] para elidir sua responsabilidade civil, comprovar que o **fato** derivou da culpa do cliente ou da força maior ou caso fortuito [...]	1.0 dano (damage) 0.25 ocorrência (event) 0.25 ocorreu (occur)	1.0 dano (damage) 0.33 lei (law) 0.22 fato (fact)
PROCESSO CIVIL. BANCO CENTRAL DO BRASIL [...] **Lei** 8.112/90 à hipótese em tela pois [...] o **crédito** em discussão decorre da relação trabalhista que outrora existia entre os litigantes [...] podendo ocasionar sérios embaraços ao orçamento do agravante, a caracterizar o fundado receio de **dano**. 9. Não há violação ao **direito** constitucionalmente assegurado de acesso ao Judiciário, eis que o recorrente poderá ajuizar ação ordinária, sendo-lhe vedado tão-somente **constituir** Certidão da Dívida Ativa [...]	1.0 dano (damage) 0.33 direito (right) 0.25 constituir (constitute)	1.0 dano (damage) 0.33 lei (law) 0.28 crédito (credit)
ADMINISTRATIVO. CADERNETA DE POUPANÇA. CORREÇÃO MONETÁRIA [...] I- A competência da Justiça Federal in ratione personae encontra-se disposta no art. 109, inciso I, da **Lei** Fundamental. [...] Reverência ao princípio constitucional da irretroatividade da **lei** para **prejudicar** o **direito** adquirido e ato jurídico perfeito [...] em respeito ao **direito** adquirido e ao ato jurídico perfeito, não calhando a alegação de negativa de vigência do art. 17 da **Lei** no. 7.730/89. [...]	0.5 prejudiciar (impair) 0.33 direito (right) 0.33 direito (right)	0.33 lei (law) 0.33 lei (law) 0.33 lei (law)
PROCESSO CIVIL - CRUZADOS BLOQUEADOS [...] POR FORÇA DA **LEI** 8024/90 [...] **RESPONDEREM** PELA CORREÇÃO MONETÁRIA [...] AS QUAIS FORAM PRIVADAS DA DISPONIBILIDADE DO **DINHEIRO** [...] NORMA POSTERIOR QUE ALTERE O ÍNDICE DE **CORREÇÃO INCIDENTE** SOBRE TAL MODALIDADE DE INVESTIMENTO [...] RECURSO DO BANCO CENTRAL IMPROVIDO E RECURSO DO BANCO DO BRASIL **PARCIALMENTE** PROVIDO.	0.5 correção (change) 0.25 incidente (incident) 0.2 responderem (respond)	0.33 lei (law) 0.26 dinheiro (money) 0.23 parcialmente (partially)

Table 1. Evaluation of the approaches on a real user case

We remarked on two important limits: i- concerning semantic similarities, since lemma maps to one or more word senses, the original translated noun *damage* is considered as a verb scoring a similarity of 0.5 due to the order: change#v#1$<^h$damage#v#1$<^h$impair#v#1; ii- concerning word embeddings the word *lei* (law) is systematically recognized as the most similar after *dano* due to the relatively smaller set of vectors obtained from the Brazilian penal code.

2.4. Interface

An illustration of the interface for the application of the framework `LooPings` [11] to display the semantic similarities is presented in Figure 2 for the `PLR` measure. The scores are finally placed on a segment $[0, 1]$. Note that because sets of answers can have the same scores (cases of ex æquo), the interface randomly choose one ID to display beside groups of points with the same scores.

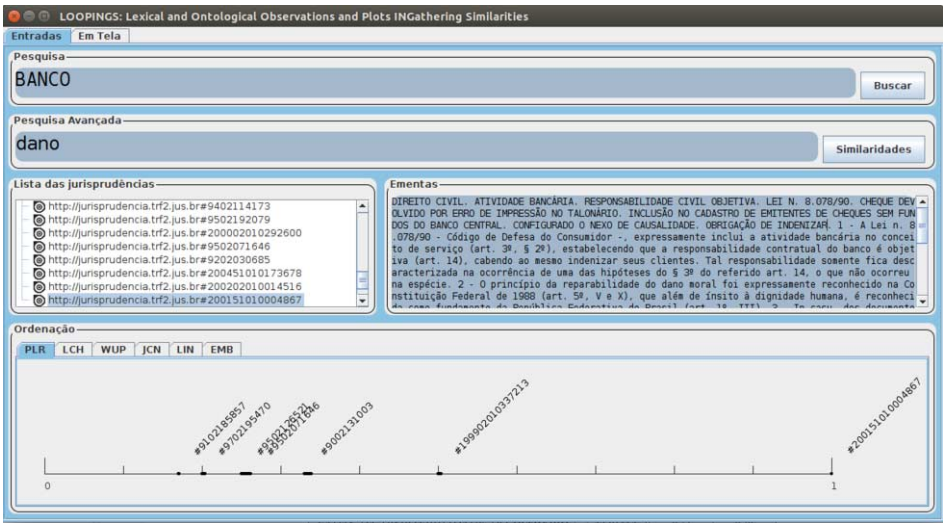


Figure 2. An interface showing the application of `LooPings`

3. Related works

The issue of ordering legal documents has already been investigated by Lu and Conrad [12] who proposed an issue-based content recommendation system with a built-in topic detection/segmentation algorithm for the legal domain.

The usage of conceptual layers to support searches among jurisprudences is a relative new trend of research in Brazil (see [13] for example). The syntactic similarities can also be used to assist such tasks because their computations can be transposable for the Portuguese language (see [14] for example), but very few works dealt directly with semantic similarities using a Portuguese material. One remarkable work was carried out by Aleixo and Pardo [15] in which node-length path similarities (after lemmatization and stop list treatments) are performed using a Brazilian Portuguese thesaurus in order to compute relatedness between sentences. In keeping with this trend, Baldez de Freitas et al. [16] proposed a measure extending the path length based relatedness in order to compute similarities between terms of distinct ontologies.

The usage of word embeddings to browse legal items is also relatively new. Landthaler et al. [17] recently explored a method that provided semantically similar answers for arbitrary length search queries using word embeddings.

4. Conclusion

In this paper, we described both ontological and non ontological approaches to perform similarities among judicial syllabi from a set of jurisprudences of a regional federal court in Brazil. We also proposed an interface to display the results of a full-text search. We now intend to propose an approach to browse the jurisprudences integrating NLP-based APIs and thesaurus in Portuguese.

References

[1] Christiane Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.
[2] Koji Sekiguchi. word2vec for Lucene, 2016. http://goo.gl/dgTxiz.
[3] Orri Erling and Ivan Mikhailov. RDF support in the virtuoso DBMS. In *Networked Knowledge-Networked Media*, pages 7–24. Springer, 2009.
[4] Giovanni Sartor, Pompeu Casanovas, Mariangela Biasiotti, and Meritxell Fernández-Barrera. Approaches to legal ontologies: Theories, domains, methodologies. law. *Governance and Technology series. Springer*, 2011.
[5] Hugo Gonçalo Oliveira. The creation of Onto.PT: a wordnet-like lexical ontology for Portuguese. In *Proceedings of 11th International Conference of Computational Processing of the Portuguese Language*, volume 8775 of *LNCS*, pages 161–169. Springer, 2014.
[6] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An open Brazilian Wordnet for Reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, 2012. The COLING 2012 Organizing Committee.
[7] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet: Similarity - measuring the relatedness of concepts. In *Proceedings of the 19th National Conference on Artifical Intelligence*, pages 1024–1025. AAAI Press, 2004.
[8] Hideki Shima. Wordnet Similarity For Java (WS4J), 2015. http://goo.gl/FTAU52.
[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
[10] Erick Galani Maziero. *Análise retórica com base em grande quantidade de dados*. PhD thesis, Universidade de São Paulo, 2016.
[11] Adama Sow and Jean-Rémi Bourguet. LooPings: A look at semantic similarities. In Delgado Y.H. Leiva Mederos A.A., editor, *Proceedings of the 2nd International Workshop on Semantic Web*, volume 1797 of *CEUR Workshop Proceedings*, pages 23–32, 2016.
[12] Qiang Lu and Jack G. Conrad. Bringing order to legal documents - an issue-based recommendation system via cluster association. In Joaquim Filipe and Jan L. G. Dietz, editors, *Proceedings of KEOD 2012*, pages 76–88. SciTePress, 2012.
[13] Rafael Brito de Oliveira and Renata Wassermann. Utilização de ontologia para busca em base de dados de acórdãos do STF. In Mara Abel, Sandro Rama Fiorini, and Christiano Pessanha, editors, *Proceedings of the IX Seminar on Ontology Research in Brazil*, volume 1908 of *CEUR Workshop Proceedings*, pages 147–157. CEUR-WS.org, 2017.
[14] Eloize Rossi Marques Seno and Maria das Graças Volpe Nunes. Some experiments on clustering similar sentences of texts in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 133–142. Springer, 2008.
[15] Priscila Aleixo and Thiago Alexandre Salgueiro Pardo. Finding related sentences in multiple documents for multidocument discourse parsing of brazilian portuguese texts. In *Proceedings of WebMedia08*, pages 298–303. ACM, 2008.
[16] Juliano Baldez de Freitas, Vera Lúcia Strube de Lima, and Josiane Fontoura dos Anjos Brandolt. Semantic similarity, ontologies and the portuguese language: A close look at the subject. In *Proceedings of PROPOR08*, pages 61–70. Springer, 2008.
[17] Jörg Landthaler, Bernhard Waltl, Patrick Holl, and Florian Matthes. Extending full text search for legal document collections using word embeddings. In Floris Bex and Serena Villata, editors, *Legal Knowledge and Information Systems - JURIX 2016*, volume 294 of *Frontiers in Artificial Intelligence and Applications*, pages 73–82. IOS Press, 2016.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-125

A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases

Isar NEJADGHOLI [a,1], Renaud BOUGUENG [a] and Samuel WITHERSPOON [a]
[a] *IMRSV Research Lab, Miralaw Inc., Ottawa, Canada.*

Abstract.
A semi-supervised approach was introduced to develop a semantic search system, capable of finding legal cases whose fact-asserting sentences are similar to a given query, in a large legal corpus. First, an unsupervised word embedding model learns the meaning of legal words from a large immigration law corpus. Then this knowledge is used to initiate the training of a fact detecting classifier with a small set of annotated legal cases. We achieved 90% accuracy in detecting fact sentences, where only 150 annotated documents were available. The hidden layer of the trained classifier is used to vectorize sentences and calculate cosine similarity between fact-asserting sentences and the given queries. We reached 78% mean average precision score in searching semantically similar sentences.

Keywords. semantic modeling, automatic annotation, semantic similarity search

1. Introduction

Systemic barriers prevent some Canadians from having adequate access to the legal system and a growing number of Canadians represent themselves in court because of the high cost of retaining a lawyer [1].

In this work, we proposed an immigration-specific search algorithm to make legal research more efficient, thorough, and user-friendly. Search engines available in Canada today are not always effective because they merely match keywords to their results and require users to use refining tools to their searches.

In our approach, which is semantic search, the meaning of words and similarity Semantic search will allow users to input natural language queries without the need to be familiar with the jargon used in legal documents and will also respond to queries semantically which includes synonyms and relevant concepts besides exact matches.

Moreover, we designed this system to find sentences that assert a fact of the case and limit the search to only these sentences. The greater the similarity between the facts of any two cases, the more likely the legal outcome, or judgment, will be similar. Thus, older cases can be used to predict new cases. By identifying fact sentences, and com-

[1]Corresponding Author: The Head of Machine Learning Research, IMRSV Research Lab, Miralaw Inc., 100 Sparks, Ottawa, Canada; E-mail: isar@miralaw.ca

paring input queries exclusively to other fact sentences, we believed we could increase the predictive accuracy of our results. Matching fact sentences with different sentence types, such as a sentence which demonstrates reasoning, or a sentence where the litigants state their positions could lead to misleading results. For example, a court may discuss a hypothetical situation to illustrate some auxiliary point. This sentence would have little to do with the facts of the case itself, and would be a poor predictor of legal outcome. Matching these sentences, which are of different types, would be misleading, because, although they are good matches, their ranking would not correlate with the cases' predictive ability, which after all, is the purpose of most legal research.

In this work, we use embedding models to capture not only the semantic meaning of words, but to model the meaning of variable-length word sequences, such as sentences, phrases or combination of keywords. We use a large corpus of immigration law cases to capture the meaning of words in the context of immigration law. This knowledge is used to train a fact-detecting classifier in a supervised manner with a relatively small set of annotated sentences. The resulted model is shown to be able to detect fact-asserting sentences of the whole corpus and the feasibility of semantic search is exhibited.

This paper is organized as following. Section 2 reviews the previous works that are related to our research problem. Section 3 explains the general approach that has been taken in this work to identify fact-giving sentences and search similar sentences to a given query. Section 4 describes the Canadian immigration law corpus that has been used in this work as well as the annotation process. In Section 5, we highlight the details of the models and training methods that have been applied in this work. Section 6 explains our evaluation methods and shows the obtained results. Section 7 summarizes the work and discusses the advantages and limitations of the proposed method.

2. Background

The idea of computer-assisted search for legal cases goes back to 1970s when Lexis legal search and Westlaw were introduced to the public [2]. Traditional legal search is based on finding exact matches to a given combination of keyword queries in a corpus of legal cases. With relentless improvement of software, hardware and natural Language processing techniques, many research efforts have been put to improve the efficiency and accuracy of finding the relevant legal cases and evaluating the results. Extracting different characteristics from a set of legal cases and weighing these characteristics was one of the approaches to improve legal search [3]. With improved calculation capabilities high-dimensional numeric features such as term-frequencies were used to represent legal texts. For example q-grams have been used to calculate similarities between investment treaties [4]. Citation analysis was one of the other advancements that was applied to improve legal information retrieval, by measuring the strength of a case based on how much other cases rely on it [5,6].

Despite all the improvements in keyword search methods, change from the use of keywords to semantics is a recent trend in legal search systems [2,7]. Finding relevant linguistics and semantic patterns has been applied for similarity search among vaccine injury decisions and was a successful step towards semantic search [8]. With the rise of word embedding models as the state-of-the-art semantic representation of words and texts, they have been used by researchers to improve search and navigation of legal data.

As examples, word embeddings have been used for detecting evidence for claims [9], argumentation mining [10] and full-text legal search [11]. Word embedding models are designed to represent a word through its contexts. In this approach, words are described in a dense and low-dimensional vector space in comparison to statistical representations of documents which describe each document as a term-frequency vector. Also, words that appear in the same context will be represented with similar vectors in word embedding models [12]. Moreover, it is impossible to represent Out-Of-Vocabulary (OOV) words in terms of term-frequency vectors [13]. However, character level word embeddings are able to assign vectors to OOV words [14].

3. General Approach

This work investigates the feasibility of semantic search among fact-asserting sentences of the legal cases. One of the challenges of training a "*legal fact detecting classifier*" is the need for adequate training data consisting of a set of legal cases annotated at the sentence level. Such an annotation process can be very costly since it can be done best by expert lawyers. We take a semi-supervised approach and show the feasibility of detecting fact sentences when only 150 annotated cases are available. Although our classifier is trained in a supervised manner, its training is initiated with the word embedding model that is trained on a large corpus in an unsupervised way.

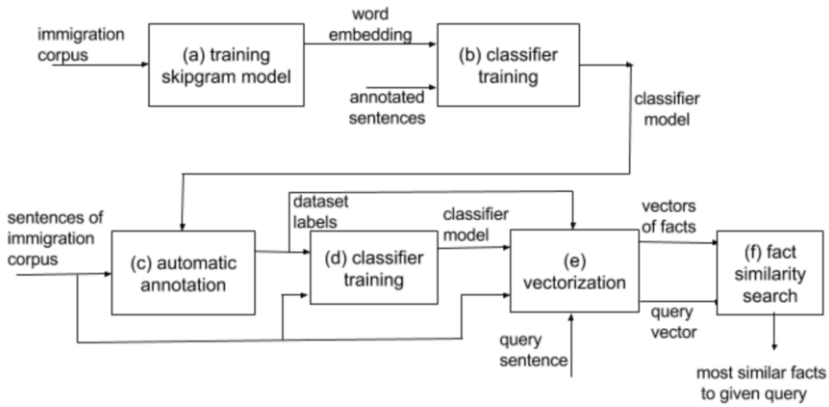


Figure 1. Steps of the proposed semi-supervised method

Figure 1 shows different steps of the proposed semi-supervised method. First, the skip-gram model [15] is trained using a large corpus of immigration cases to capture the meaning of words based on their context in immigration cases (block (a)). A small set of annotated sentences are then used to train a binary classifier that is able to distinguish between facts and non-facts (block (b)). The immigration word embedding is used as initial word representation for classifier training. The trained classifier is then used to automatically label all the sentences of immigration corpus as facts and non-facts (block (c)). This classifier is a shallow neural network with one hidden layer, the values of which can be used as vector representations of sentences in order to calculate sentence similarity. The network is fully connected and the number of neurons of hidden layer equals to the

dimension of vector space which is 100 in this work. The classifier is re-trained with the sentences of automatically annotated immigration corpus in order to capture the meaning of all words of the vocabulary and improve sentence vector representations (block (d)). The re-trained classifier is then used to get the vectors of all the fact sentences of the corpus (block (e)). For a given query, the hidden layer of the trained classifier is used to calculate the vector representation of the query (block (e)). The similarity between the query and the fact sentences of the corpus is then calculated and the most similar facts to the query are returned to the user along with a link to each of the corresponding legal cases (block (f)).

4. Dataset

We use a dataset of 46000 immigration and refugee cases available on Canada's Federal and Supreme Court websites. The HTML documents are first converted to text. Most of the documents contain headers and footers which provides specific information about the case such as date, case name, etc. Documents are processed and headers and footers are removed. The documents are then parsed to sentences. Sentences with less than 20 characters or more than 1000 characters are removed as most of them are a result of wrong sentence spiting and are only 1% of the sentences. We used the Spacy package for sentence splitting and created a set of rules to improve the quality of the sentences splitting considering the specific structure of legal documents such as paragraph numbering, titles, etc. Sentences were tokenized and the punctuations were removed. The cleaned corpus contains more than 136M words, 4549809 sentences and vocabulary size of 125846.

Table 1. Annotation Scheme

Tag	Freq.	Description
Procedure	1%	The nature of the case which is the description of the appeal and the case's procedural history and how the case was treated at previous court levels and/or tribunals.
Fact	46%	The applicant's background information, the applicant's account of his or her story, and the findings made by the tribunal member or previous judge. In the context of immigration this encompasses everything that happened in the administrative tribunal.
Party Position	13%	The applicant and respondent's respective arguments, what they were seeking, and their interpretations of the facts.
Issue	2%	The legal questions the judge must answer/decide upon including the issues ultimately not answered.
Analysis	28%	The judge's decision making process, why and how the judge came to his or her conclusions including any reference to previously decided cases.
Conclusion	6%	The sentences that provide the judge's answer to the issues.
Judgment for Appellant	1%	Statements indicating that the judge decided in favor of the applicant including both orders & holdings.
Judgment for Respondent	1%	Statements indicating that the judge decided in favor of the applicant including both orders & holdings.

4.1. Manual Annotation of Sentences

Two law students manually parsed 150 random cases (each annotator 75 cases) to sentences and annotated the sentences using eight different sentence tags. The detailed de-

scription of the tags as well as their frequency in the 150 annotated cases are provided in Table 1. Sentences that did not fit in any of these tags remained unannotated. The sentences that could viably correspond to more than one potential tag, were tagged according to their most dominant characteristics. From Table 1, we can observe that about half of the sentences are tagged as fact. Therefore, this dataset is balanced for training a binary classifier to detect facts.

5. Training Methods and Procedures

5.1. Unsupervised Semantic Modeling of Legal Words

Distributed word representations or word embeddings are introduced to capture the semantic meaning of words by assigning vectors to each word. Word embeddings have been vastly studied and used in NLP applications [15,16,17,18]. One popular example of building word embeddings is the skip-gram model introduced in [15], where the distributed representations are trained to predict words that appear as their neighbors in the training corpus. The objective function to be maximized during training is:

$$\sum_{t=1}^{M} \sum_{i\neq t, i=t-N}^{t+N} \log p(v_{di}|v_{dt}) \tag{1}$$

summed across all words v_{dt} in all documents v_d, where M is the number of words and N is the length of skip-gram window and p is the probability of occurrence of v_{di} as a neighbor of v_{dt}. We trained a skip-gram model using the dataset described in Section 4 to build a word embedding specialized in immigration law. The dimension of embedding vectors is 100. The embedding features are word n-grams where $n = 1, ..., 4$.

Although pretrained word embeddings are available and provided by NLP tools, for semantic search in immigration legal corpus, a word embedding model trained on the same corpus is preferred because it captures the legal meaning of terms. For example, in an immigration law word embedding, the word ***immigration*** is found to be close to word ***FCJ***, which is a frequently occurring Federal Court citation component and ***IRPA***, which is the short form of the *Immigration and Refugee Protection Act*. However, in a general word embedding (provided by Spacy library), ***immigration*** is closest to general terms such as ***reform*** or ***citizenship***. This is because our legal word embedding is trained to convey specific legal meaning of the words in the context of Canadian immigration law whereas the general word embeddings trained on general corpora carries the general meaning of immigration. Another example is the term ***allowed*** which locates near words ***dismissed, costs, ordered, assessment*** in immigration law word embedding and near words ***allow, allowing, not, unless*** in general word embedding provided by Spacy. Among different tools that are available for training a skip-gram embedding space, we chose fastText; developed and implemented by Facebook research team [14]. FastText is very fast in training in comparison to other implementations of skip-gram model and achieves almost the same accuracy. More importantly, it provides vectors for OOV words, since it is trained in character level and uses character n-grams to calculate word vectors.

The quality of a word embedding model is often evaluated in calculating word analogies besides finding most similar words to a query. Table 2 shows some of the interesting

analogies that were produced by trained word embedding model and shows examples of similar pairs of words in the context of Canadian immigration law. Word pairs (w_1, w_2) and (w_3, w_4) are shown in the same row of this table, if $w_1 - w_2 = w_3 - w_4$, where w_i, represents a word vector. [2]

Table 2. Analogues word pairs found from Canadian immigration law word embedding.

Pair1	Pair2
China - Chinese	Sri Lanka - Sri Lankan
Colombian - FARC	Somalian -Alshabab
Roma - Hungarian	Bahai - Iranian
Palestine - Hamas	Lebanon - Hezbollah
PRRA - Preremoval	RPD - posthearing

5.2. Automatic detection of fact-asserting sentences through supervised learning

We used the annotated dataset, described in Section 4.1, to train a binary classifier that automatically detects fact-asserting sentences in immigration cases. We used the supervised model from the fastText library to achieve this classification task. Figure 2 shows the model architecture of fastText supervised classifier for a sentence with N n-gram features, $w_1, \ldots, w_N$ [19] where w_i is the embedding vectors corresponding to i^{th} feature. In this model, the text representation is a hidden variable which can be potentially used as a text representation in other tasks. This architecture is similar to the Continuous Bag of Words (CBOW) model of [15], where the middle word is replaced by a label. The hidden variable is then mapped to the class label through a softmax output layer with the number of neurons equal to the number of classes. The softmax function is used to compute the probability distribution over the predefined classes. For a set of D documents,the following negative likelihood function is minimized.

$$-\frac{1}{D}\sum_{d=1}^{D} y_d \log(f(UWv_d) \tag{2}$$

where v_d is the normalized vector representation of the d^{th} document, y_d represents the label, U and W are the weight matrices. We trained this model on a 4-core CPU machine. The cost function is optimized using stochastic gradient descent and a learning rate that decreases linearly. We used the trained embedding model, described in section 5.1, to initialize the embedding layer of this classifier. In theory, the embedding layer of this model can be initialized either with random vectors or a pre-trained word embedding. we only have 12220 annotated sentences (from 150 cases), which might not be enough to train both embeddings and classifier layer of the structure shown in Figure 2. Therefore, we initialize the first layer with the embedding vectors trained on the immigration law corpus. The trained classifier is used to automatically tag all the sentences of the corpus as facts or non-facts.

[2] In Table 2, PRRA is an acronym for Pre-removal Risk Assessment which is an unsuccessful refugee claimant's last chance to avoid deportation. RPD (Refugee Protection Division) is the body that hears Refugee matters in Canada and posthearing is the descriptor of the consequences of applying.

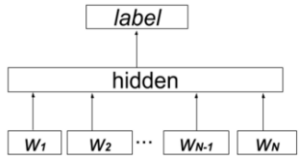


Figure 2. fastText supervised classification model

5.3. Database vectorization and similarity search

Although the classifier explained in Section 5.2 is primarily designed and applied for classification of short documents such as sentences, the values of the hidden layer of the trained model can be used to calculate a vector representation for a given sentence. However, the size of the vocabulary of the annotated documents is much less than the size of the vocabulary of the whole corpus. Therefore only a small portion of word vectors are updated during training of the classifier. To improve the effectiveness of sentence vectorization, we use the automatically annotated sentences of the legal corpus to retrain the classifier. In this way, we update the word vectors of the classifier for all the words of the corpus. After re-training of the classifier, the values of its hidden layer are taken as vector representation of all the sentences present in the corpus. These vectors are stored for similarity search purpose only if the assigned label to the sentence is fact.

For a given query, which can be a sentence, a phrase or a combination of keywords, the re-trained classifier is used to label and vectorize it. We compute the cosine similarity between the query vector and vector of each fact sentence in the corpus. The resulting similarity scores are sorted in a descending order and the three most similar sentences are found and their corresponding documents are returned. The intuition is that sentences which are most similar to the query sentence should rank at the top of the retrieval results. The similarity score can also be understood as an estimate of the relevance of a sentence with respect to the query.

6. Evaluation and results

6.1. Evaluation of the classifier

Although the classifier was trained using sentences parsed by human, we valuated the trained classifier on sentences that were parsed automatically, since that is the ultimate performance of the classifier that the user experiences. 300 automatically parsed sentences were randomly selected from the corpus (annotated documents described in section 4.1 were excluded). These 300 sentences were manually annotated by experts to be used as the test set for classification. In this test dataset, 47% of the sentences are stating a fact and the rest are tagged as non-facts.

In order to compare the classification accuracy of the classifier described in Section 5.2, with benchmark classifiers, we trained six different binary classifiers to detect fact sentences using the training dataset described in Section 4.1 and tested using the test set. The description of theses classifiers as well as the obtained classification accuracies for the test set are given in Table 3. These results show that the proposed semi-supervised method outperforms commonly used classifiers in detecting facts when the amount of training data is relatively small.

Table 3. Classification results for detecting fact sentences using various binary classifiers.

Classifier Description	Acc.
Sentences are represented with *term frequency-inverse document frequency* (tfidf) features. A SVM binary classifier is trained.	81%
Sentences are represented by averaging of word vectors from the embedding space trained in Section 5.1. A SVM binary classifier is trained.	83%
Sentences are represented by tfidf weighted averaging of word vectors from the embedding space trained in Section 5.1. A SVM binary classifier is trained.	84%
A fastText supervised model is trained with random initial word embeddings.	83%
A fastText supervised model is trained with pre-trained word vectors (provided by fasteText) [20] as initial values of embedding layer.	86%
A fastText supervised model is trained with immigration law word vectors (described in Section 5.1) as initial values of embedding layer (proposed semi-supervised method).	90%

6.2. Evaluation of the proposed similarity score

The goal of this evaluation is to measure the algorithm's ability to return semantically relevant sentences as top results, given a query.

We used the Mean Average Precision (MAP) metric, which is a standard comparative evaluation metric for search engines [21] and indicates how precisely the relevant sentences can be ranked on top, in a set of candidate sentences, based on their similarity score to the query.

Table 4. Candidate sentences, corresponding human judgments (HJ) and calculated similarity scores (SS) for the query *"Applicants PRRA rejected despite his fear of persecution and violence in Sri Lanka."*

Sentence	HJ	SS
He claimed to have a well-founded fear of persecution and argued that Sri Lanka was violent, nevertheless, the PRRA rejected the application.	R	0.96
The PRRA Officer reviewed the Applicants immigration history and quoted extensively from his statutory declaration dated March 26, 2014 including the Applicants claim that, even though the war in Sri Lanka has ended, the situation there is worsening in many ways and that, if returned, he would face discrimination and harassment due to his ethnicity and would be targeted because he has family overseas.	R	0.93
The applicant is a member of Tamil and claims fear of persecution.	R	0.85
With respect to the male Applicant's claim, the Board held that he did not have a well-founded fear of persecution because he was not really wanted by the Iranian authorities.	N	0.74
Specifically, the Applicant argues the RPD erred by: making plausibility findings without specific reference to the evidence to support such findings; making an overall credibility finding before independently assessing his corroborative evidence; discounting the psychiatrists report; and, failing to address his claim that he was kidnapped by authorities in 2013.	N	0.12

We designed an evaluation dataset with human judgments on semantic similarity. The evaluation dataset is a collection of 15 queries, crafted by legal experts, each targeting one important area within Canadian Immigration Law. For each query, a set of 5 candidate sentences was built which was a mix of sentences handpicked from the CanLii website [22] and sentences handcrafted by the evaluators. The evaluators assessed the relevance of a sentence with respect to its query by marking it as "Relevant" or "Not Relevant". The candidate sentences were ranked based on their similarity score to the query and Average Precision (AP) was calculated for each query to measure how precisely the

"Relevant" candidates are ranked higher than "Not Relevant" candidates. As an example, Table 4 shows the candidate sentences, human assessments and similarity scores for the query *"Applicants PRRA rejected despite his fear of persecution and violence in Sri Lanka."*. We simply calculated the mean of all APs over all queries and obtained the MAP score of 78%.

7. Discussion, limitations and scope of use

We showed the feasibility of detecting fact-asserting sentences and searching for semantically similar facts in a large Canadian immigration law corpus when only 0.3% of the corpus is manually annotated. Figure 3 shows a screenshot of a system developed based on the proposed method.

Our evaluation show that a supervised fastText classifier that is initiated with immigration law word embedding is more effective than benchmark classifiers in identifying fact sentences (see Table 3). Evaluation of the proposed semantic similarity score has been carried out using a small hand-crafted evaluation dataset and an acceptable MAP score is acquired. The main advantage of the proposed similarity score is that it automatically limits the search to facts given a fact query, since the similarity score between a fact sentence and a non-fact sentence calculated by the proposed method is very low even if the two sentences share semantically similar words. The other advantage of this method is that slight misspelling of words does not change the results, since fastText is a character level word embedding. An alternative method of semantic score calculation such as tfidf weighting of word vectors will completely ignore misspelled words, since they are not included in the vocabulary of tfidf calculator. A more rigorous quantitative evaluation of this search method and comparing it with other alternatives remains as a focus of future work due to challenges of designing a comprehensive evaluation dataset.

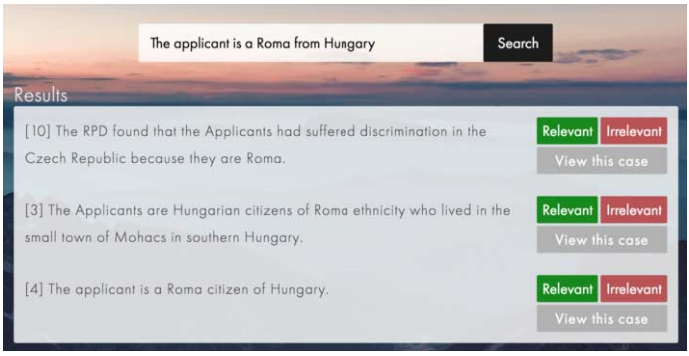


Figure 3. Example of the results returned by the developed system.

Acknowledgement

The authors would like to thank Ms. Chelsea Kirsch and Mr. Micheal Elharrar for annotation, their enormous help in formulating the research question in the context of Canadian legal environment and also evaluating the results. This research was made possible in part with funding from the Canadian Industrial Research Assistance Program (IRAP).

References

[1] R. Birnbaum, N. Bala, and L. Bertrand, The rise of self-representation in canadas family courts: The complex picture revealed in surveys of judges, lawyers and litigants, Canadian Bar Review **91**, 2013.

[2] J. O. McGinnis and R. G. Pearce, The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services, Fordham L. Rev. **82**, (2014), 3041-3050.

[3] Q. Lu and J. G. Conrad, Next generation legal search-its already here, Vox Populii blog, Legal Information Institute (LII), Cornell University, 2013.

[4] W. Alschner and D. Skougarevskiy, Consistency and legal innovation in the bit universe, Stanford Public Law Working Paper No. 2595288, 2015.

[5] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck, Network analysis and the law: Measuring the legal importance of precedents at the us supreme court, Political Analysis **15** (2007), 324346.

[6] R. Winkels, A. Boer, B. Vredebregt, and A. van SOMEREN, Towards a legal recommender system, JURIX **271** (2014), 169178.

[7] E. Francesconi, S. Montemagni,W. Peters, and D. Tiscornia, Semantic processing of legal texts: Where the language of law meets the law of language, Lecture Notes in Computer Science **6036**, Springer, 2010.

[8] M. Grabmair, K. D. Ashley, R. Chen, P. Sureshkumar, C. Wang, E. Nyberg, and V. R. Walker, Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools, ACM (2015), 6978.

[9] R. Rinott, L. Dankin, C. A. Perez, M. M. Khapra, E. Aharoni, and N. Slonim, Show me your evidence: an automatic method for context dependent evidence detection, EMNLP (2015), 440450.

[10] N. Naderi and G. Hirst, Argumentation mining in parliamentary discourse, International Workshop on Empathic Computing (2014), 1625.

[11] J. Landthaler, B.Waltl, P. Holl, and F. Matthes, Extending full text search for legal document collections using word embeddings. JURIX (2016), 7382.

[12] K. Erk, Vector space models of word meaning and phrase meaning: A survey, Language and Linguistics Compass **6** (2012), 635653.

[13] F. Huang and A. Yates, Distributional representations for handling sparsity in supervised sequence labeling, ACL-AFNLP **1**, Association for Computational Linguistics (2009), 495503.

[14] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics **5**, (2017), 135146.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.

[16] D. E. Rumelhart, G. E. Hinton, R. J. Williams, et al., Learning representations by back-propagating errors, Cognitive modeling **5**, 1988.

[17] F. Morin and Y. Bengio, Hierarchical probabilistic neural network language model, AIStats **5** (2005), 246252.

[18] J. Pennington, R. Socher, and C. D. Manning, Glove: Global vectors for word representation., EMNLP **14** (2014), 15321543.

[19] A. Joulin, E. Grave, and P. B. T. Mikolov, Bag of tricks for efficient text classification, EACL (2017), 427-430.

[20] Pretrained fasttext word vectors. *https://github.com/facebookresearch/fastText/blob/master/pretrainedvectors.md*, Accessed at 24-08-2017.

[21] A. Turpin and F. Scholer, User performance versus precision measures for simple search tasks, ACM (2006), 1118.

[22] Canlii website, *https://canlii.org*, Accessed: 2017-06-30.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-135

Toward Building a Legal Knowledge-Base of Chinese Judicial Documents for Large-Scale Analytics

Amarnath GUPTA [a,1], Alice Z. WANG [b], Kai LIN [a], Haoshen HONG [a], Haoran SUN [a], Benjamin L. LIEBMAN [b], Rachel E. STERN [c], Subhasis DASGUPTA [a], and Margaret E. ROBERTS [a,2]

[a] *University of California San Diego, USA*
[b] *Columbia Law School, USA*
[c] *University of California Berkeley, USA*

Abstract. We present an approach for constructing a legal knowledge-base that is sufficiently scalable to allow for large-scale corpus-level analyses. We do this by creating a polymorphic knowledge representation that includes hybrid ontologies, semistructured representations of sentences, and unsupervised statistical extraction of topics. We apply our approach to over one million judicial decision documents from Henan, China. Our knowledge-base allows us to make corpus-level queries that enable discovery, retrieval, and legal pattern analysis that shed new light on everyday law in China.

Keywords. legal ontology, information extraction, knowledge representation, topic model, Chinese legal documents, text analytics

1. Introduction

In recent years, governments around the world have moved to make information about their legal systems more transparent in order to hold courts accountable to the public and inform legal participants of past court behavior. In Europe, the `OPENLAWS.eu` Consortium is developing an open platform where laws, cases and legal literature from all member states will be made publicly available [24]. In China, the court system recently began mandating that courts upload decision documents to the public website run by the Supreme People's Court (SPC) [17]. While millions of documents are available in each of these contexts, much of the information in the documents is unstructured, and therefore not useful in aggregate for the public. As larger and increasingly more complete collections of legal data become available, there is a corresponding need to construct *publicly available legal knowledge-bases* – formal representations of legal information – from these documents to facilitate their analysis.

The idea of creating legal knowledge-bases, and more generally knowledge-based systems, is not new [23,12,5]. Legal knowledge-bases have been developed in the past

[1]Corresponding Author, E-mail: a1gupta@ucsd.edu
[2]Corresponding Author, E-mail: meroberts@ucsd.edu

for diverse tasks like citation analysis [11], e-governance [12], criminal law analysis [8] and legal advice systems [25]. However, the scope of these tasks has mostly been confined within a single document and in some cases, to small databases: understanding the provisions of a particular law, the argumentation structure of a particular legal case, or the logical reasoning of a particular court procedure. While these analyses can be useful for specific problems, much can be gained by building knowledge-bases to support large-scale analyses that inform legal researchers about the deep characteristics of the complete collection. The goal is to enable "reading at a distance" [13], by capturing knowledge that helps a researcher uncover patterns and emerging trends that can only be mined from a large legal corpus. These analyses can also be reintegrated as a statistically derived knowledge-item into the knowledge base to be reused in subsequent analyses. We call this form of analyses *legal pattern analyses* (LPA).

In this paper, we present an approach for constructing a logically sound legal knowledge-base that allows for large-scale analyses. We apply this approach to a corpus of over 1.1 million judicial records from Henan China.[3] On this example corpus, the purpose for the knowledge-base is to enable a user perform such tasks as:

- *Knowledge-based Retrieval:* "Retrieve the most common defendants in administrative cases." Administrative litigation cases in Chinese law are those where individuals are most likely to challenge the government and therefore are of interest to political scientists studying citizen activism in China [19]. What types of government entities are the most common targets of these cases?
- *Knowledge-based Discovery:* "Discover the issues of dispute for divorce-related cases where the plaintiff is female." Women are known to be disadvantaged in divorce cases under Chinese law [10,16,14]. In what circumstances do they use the legal system to file complaints?
- *Knowledge-based Pattern Analysis:* "Calculate the major differences between cases where plaintiffs file individually versus collectively in administrative cases against the government." Collective action against government entities is viewed as politically sensitive because it could spill over into protest [6], and courts sometimes break up collective claims into individual lawsuits for this reason. On what issues is the government sued by a collection of individuals in the Chinese legal system?

Challenges. There are some inherent challenges in creating a knowledge-base that is conducive to a general set of corpus-level analyses that this paper seeks to address.
Linguistic Variability. Unlike knowledge-bases over formally-written legislation, judicial decision documents (JDDs) have a variable format. For example, arrest records of the defendant in a criminal case or decisions show wide variations in structure and level of detail. Hence, linguistic processing of JDDs for knowledge extraction is more complex, particularly when extracting the same information from the entire corpus.
Need for Heterogeneous Representation. No uniform knowledge representation technique can practically capture all requirements of the knowledge-base. This problem has been reported in prior research. For example, [1] uses description logic for facts and a logic programs for rules, while [2] uses a hybrid rule-based/case-based model for divorce dispute resolution. Our knowledge-base must satisfy retrieval, discovery, and large-scale analytics, each requiring different inputs.

[3] More information about this corpus and what it represents is in [17].

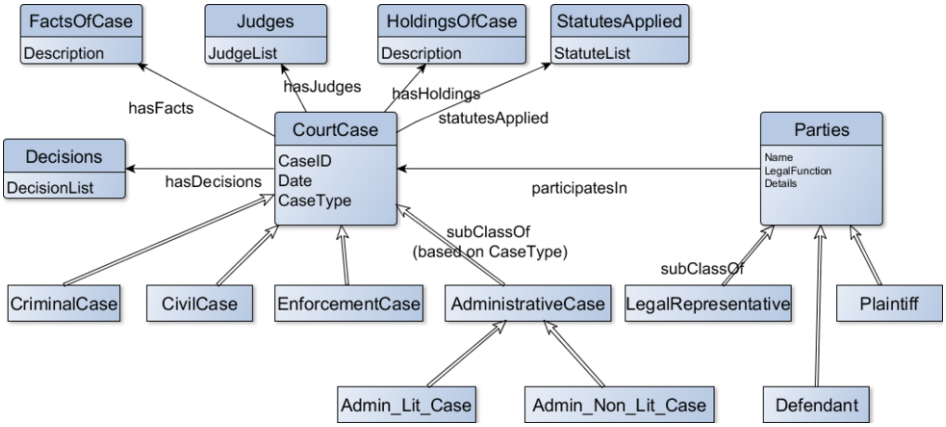


Figure 1. The basic schema of the Chinese Judicial Decision Documents. Several attributes of the CourtCase entity have been omitted for clarity. A simple arrow represents a relationship while a double-shaft arrow represents a subclass relationship. Here "admin_non_lit" stands for "Non-litigation administrative enforcement".

Lack of Completeness. "Knowledge" in the knowledge-base is always incomplete and sometimes inconsistent. Legal ontologies such as Core Legal Ontology (CLO)[4] do not capture concepts and relations that are in the documents but have not been formalized. The challenge is to be able to relate unstructured elements of the text to a legal ontology.

2. The Anatomy of a Chinese Judicial Decision Document

A JDD is written as unstructured text; however, because legal texts are formulaic, sections of the texts can be parsed. We take a data-centric approach to the problem [4] and apply an initial parsing [17] to extract a roughly relational structure, whose extended entity-relationship diagram [9] is shown in Figure 1.[5] This structure is rough because the exact content of a JDD depends on the case type it represents; for example, a criminal case includes information about prior criminal history, whereas an administrative case does not.

Figure 1 shows the schema of a case after initial parsing. "Parties" contains a list of individual party members, including a text description of each party, its role in the court case (e.g., plaintiff), and when applicable its relationship to other parties (e.g., the guardian of a minor defendant). The "factList" contains unstructured text of the legal facts of the case, a summary of primary arguments by the two sides as well, and the facts as established by the court. The "holdingList" contains the legal reasoning and analysis of the judge, which applies the law to the facts of the case. The "decisionList" specifies the legal verdict of the court including any judgments or case-dismissal statements. Both court cases and parties have subtypes. We only show some of the subtypes in Figure 1, and point out that while the subtypes of a court case can be syntactically recognized from the case identifier, the subtypes of the parties can be recognized only through text processing (see Section 3.2).

[4] `www.loa.istc.cnr.it/ontologies/CLO/`

[5] All analysis is done in the original language of the legal documents, Chinese.

Schema Element	Ontology Concept
CourtCase	'Legal case' `CLO:CoreLegal.owl#LegalCase`
CourtCase.caseType	`subClassOf LegalCase`
Parties	`is-participant-in some CLO:CoreLegal.owl #LegalFunction`
Holding	`LegalAnalysisDescription` ⊓ `analyzedBy some Court`
Decision	'Judicial Decision' `CLO:CoreLegal.owl#JudicialDecision`
Fact	'Legal fact' `CLO:CoreLegal.owl#LegalFact`
Statute	'Law' `rdf:type CLO:CoreLegal.owl#Law`
Judge	`dbpedia.org/ontology/Judge`

Table 1. Ontology to Schema mapping in our knowledge-base

3. Our Approach

To build the knowledge-base, we take the following approach. We start with an existing initial ontology, which, although incomplete, maps well to the basic EER diagram in Figure 1. We adjust this ontology to ensure alignment with the all elements of the EER schema. Next, we extract semi-structured information from the JDDs with two different techniques. Then, we conduct a two-way annotation process from the ontology to the JDDs and from the JDDs back to the ontology. The annotated ontology and documents are stored in a scalable polystore system [7]. Finally, we compute a family of topic models on the data to create statistical representations of the remaining unstructured text.

3.1. An Initial Ontology

Our initial ontology is derived from two well-known ontologies in the domain of legal knowledge representation. The upper ontology is DOLCE+DnS Ultralite (DUL) ontology,[6] which was chosen primarily because of its elaborate coverage of the concept space including `social objects`, `Conceptual Objects` (called `concepts`) and `situations`. The domain ontology is adapted from the Core Legal Ontology (CLO) which, in turn builds on DUL and the Information Object Ontology Lite.[7] The CLO introduces the basic concepts of jurisprudence including `law, legalFunction, legalDescription, crime` and `legallyRelevantCircumstance`.

Schema Alignment. The schema elements of Figure 1 are first mapped to the ontological concepts in Table 1. Next, we directly relate the `caseType` attribute to the subclasses of the `LegalCase` concept. The mapping for `Parties` implies that every party in the list of parties plays the role of a `legal function` as specified in the CLO, which designates plaintiffs, defendants, attorneys, etc. as legal functions that are fulfilled by concrete instances of `NaturalPerson` entities. Similarly, a statute is interpreted as an individual instance (`rdf:type`) of the concept of `law`. The `holding` is mapped to our extension of the CLO which admits the concept of `LegalAnalysisDescription` ⊏ DUL.`Description`. This common structure of cases can be extracted from each case fairly easily, but still draws only basic information from each decision.

Initial Ontology Augmentation. In CLO, the concept of `law` (corresponding to a statute in the schema) is the subclass of a `legal description`, which is the subclass of the

[6] `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`

[7] `www.ontologydesignpatterns.org/ont/dul/IOLite.owl`

CLO concept `description`. CLO also defines the concept of `legal case` as a derived subclass of the DUL concept of `legal fact` which depicts situations depending on legal norms. For example, the legal case called `crime` satisfies norms of incrimination. But how does a concrete crime type such as arson ("setting fire" – 放火) relate to the concept of `law`? We extend CLO by creating a subclass tree under `law`. The tree is derived from the case classification documentation issued by the Supreme People's Court in China. This subtree categorizes laws at a level of detail that can be more effectively correlated with the judicial decisions documents. For example, the `tort liability law` (侵权法) is a subclass under the concept `civil laws` (民法) branch. We also introduce a new subclass hierarchy under `legal case` to represent a hierarchy of legal case types (e.g., `Product transporter responsibility dispute`, 产品责任纠纷, is a superclass of 产品运输者责任纠纷). The rest of the ontology classes where assigned to the JDDs based on existing categories available from the SPC website.

The resulting ontology is checked for consistency with Protégé's *Hermit reasoner* and then stored in a graph database system (Neo4J) through *SciGraph*,[8] an ontology manager developed over Neo4J. SciGraph uses the OWL API to decompose each axiom and a model conversion algorithm to re-represent them as graph. The graph nodes are typed and can represent concepts, individuals and anonymous classes; the edges represent subClassOf, equivalence, union (⊔), intersection (⊓) that are used in the ontology. SciGraph is a lossless representation of the asserted ontology – its edges capture quantifiers (i.e., `some, only`, ...) and edge properties like transitivity. Simple inference procedures like transitive closure computation are implemented through graph-based operations. This implementation supports knowledge-based querying (Section 4).

3.2. Information Structuring with Text Analytics

While the ontology and its mapping to the JDD schema creates a preliminary connection between them, much of the information content of the JDD is still buried inside its unstructured content. We will describe two methods by which we extract information from text into a *semistructured* (JSON) representation. This semistructured (labeled, ordered trees) model provides an additional advantage that the extracted information can be stored in a scalable semistructured database like MongoDB.

Term-Anchored Context-free Grammar. Our first approach of knowledge extraction from text applies context-free grammar rules to segments (e.g., `Parties`) of a JDD where "anchor terms" from a large but fixed vocabulary must appear . Our intention is to extract a complex set of properties of entities mentioned in the document, and the complex relationships between these entities. To see why this is important, consider the analyst's question: "Do repeat offenders get harsher sentences?" To determine whether a defendant is a "repeat offender" one has to extract the criminal record from the description of the defendants. In our example collection of JDDs, most criminal defendant descriptions present a history of their criminal record, although these descriptions are not standardized.

We take a grammar-based approach to information extraction from text. We argue that since the description of criminal records is "stylized" natural language, its grammar falls somewhere between a pure, context-free grammar (CFG) and an arbitrary context-sensitive grammar. We postulate that if we recognize a handful of *anchor terms* in the

[8] `https://github.com/SciGraph/SciGraph`

Action Prohibition	Original Judge Affirmation	Remand
Case Withdrawal	Custody of Child	Monetary Compensation
Confirm Illegality	Punishment Announcement	Property Distribution
nolle prosequi	Judgment Revocation	Penalty Abatement
Confiscation	Compulsory Execution	Divorce Approval

Figure 2. Examples of the 38 sentence categories parsed by our sentence modeling scheme

text, then the rest of the text can indeed be treated as though it satisfies a CFG grammar. The anchor terms are identified through different dictionaries such as the dictionary of law enforcement actions and the dictionary of charges that can be brought by the police. The grammar rules are centered around terms in these dictionaries, such as "imprisonment" or "drug possession," then a context-free rule can correctly extract the prior record. A preliminary evaluation shows that parties are correctly assigned in 85% of the cases. The errors primarily occur due to complex unparsed sentences, and in documents where there is no specific party section but the case title carries the information about parties.

Judicial Sentence Models. Our second approach to information extraction relies on sentence modeling and is applied toward understanding the court decisions. The first step in this approach uses the output of the CFG party extraction described before to instantiate participant names, aliases, and their roles in the decision section of the document. A Jaccard coefficient based scoring method is used for inexact matches and abbreviations.

The second step creates a classification of the types (Figure 2) of verdict sentences through a series of matching rules. As a simple example, the sentence "Dissolving the plaintiff Xu Shouzhen's and the defendant Liu Weihong's marriage relationship" can be easily categorized as a *marriage dissolution* verdict because it has the sentential pattern `removing <plaintiff-phrase> and <defendant-phrase> marriage relationship`. Similarly, a pattern `Criminal + <name> + commit + ... + crime` classifies it as a *Punishment Announcement* verdict. The complexity of the classification rules arises from the syntactic variations in the sentence structure and the context sensitive nature of the text. The recognizer of a compensation case may use synonyms and expression variants like "to ... compensate ..RMB." (向...赔偿...元..) In other cases, the classification rules must look at multiple consecutive sentences to provide adequate context.

Once a sentence is classified into one or more of 38 classes, we reanalyze the sentence to identify model parameters. For example, a compensation case will have payer(s), a set of payee(s), and a compensation amount for each payer-payee combination. When the compensation amounts are explicitly specified, we record them; when clauses like "equally paid" are used, they are specifically interpreted to determine the actual compensation amount. Often verdicts have additional clauses such as "payable once every year by October 1" – these clauses are captured within a "comment" node in the resulting tree.

If a sentence corresponds to multiple possible models, a *conflict resolution* process is applied. For example, a verdict that affirms the original judgment always includes rejecting other requests. This verdict will be identified as both "Affirm Original Judgment" and "Reject Requests" types in model selection stage. In this case we order the verdict types by their frequency of occurrence, and select the top scoring model. Our preliminary evaluation shows that the sentence classification has over 95% accuracy for tweets

with correctly parsed parties; 5% error-rate is due to complex sentences in the decision section that could not be parsed properly.

3.3. Bidirectional Mapping

The next step is a two-way mapping from the ontology to the restructured JDDs and from the JDDs back to the ontology. The rationale for the two-way mapping comes from the observation that analysis of a JDD corpus yields new concepts, individuals and relationships that should be included in an application ontology that "hangs from" the domain ontology from Section 3.1. Simultaneously, the process creates an ontological annotation into the semistructured data that explicitly marks ontological concepts/individuals to the JDDs. For instance, the term "Entrusted Agent" is a new instance of `legalFunction`, "arrest record" is a new concept, and $civil-case \overset{describedBy(some)}{\longrightarrow} civilLaw$ is a new relationship that would be added to the ontology. In the other direction, we annotate the ontology with JDD-indexing mappings such as $attorney \overset{mapsTo}{\longrightarrow} entrustedAgent \overset{occursIn}{\longrightarrow}$ 100.*Parties*.3 where the first element `attorney` is a CLO concept, the second element `entrustedAgent` is a party type and the third element `100.Parties.3` represents the 3rd Parties element in document having ID 100. These are encoded in the JDD data as JSON element `mappedEntityType` added to every recognized instance of a concept.

3.4. Leveraging Unstructured Text

Last, we leverage unsupervised natural language processing to extract information from the remaining unstructured text. Topic models have been amply used for tasks related to legal document understanding as diverse as extracting domain and argument related words [20], legal document summarization [15], finding differences in decision patterns across courts [18] and shifts in the content of the case-law of international courts over time [21]. They identify "topics," or clusters of frequently co-occurring terms in a collection of documents [3].

In our setting, we estimate the Structural Topic Model (STM) [22] over the results of a query which subsets the data based on some conjunctive predicate P. The predicates may place conditions on metadata (date $>$ 1/1/2014), or document content (e.g., Facts.factList contains "pollution"), or on derived structures (e.g., verdict type = "Punishment Abatement") or any conjunction of the above. Further the topic model can be run on any subset of the parts of the document (e.g., only facts and decisions) – this subset is called the "scope" S of the model. This *PS conditioning* allows us to run multiple topic models on the same collection of legal documents, giving insight into the topics tailored to the analyst's interest. Each *PS* pair has a ranked topic-term list, and a ranked topic-document list. Further, if a term discovered in a topic belongs to the ontology, it is annotated by the ID of the ontology term. Ontological annotations can also be included in the estimation of the topic model by including them as covariates in the STM.

We illustrate the effect of *PS-conditioning* on the estimated topics by running a 30-topic STM on all civil cases, restricting the scope of the model to text in facts and holdings. Row 1 of Table 2 shows one interesting topic retrieved from the model related to medical care. Rows 2 and 3 show how this topic becomes more refined as increasingly restrictive predicates are added.

Predicate	Scope	Topic
(none)	facts, holding	Hospital, medical expenses, disability, compensation, care, plaintiff, calculation, cost, injury, transportation
contains(document, 'disability')	facts, holding	identification, calculation, disability, compensation, forensic, identification, hospitalization, mental
contains(document, 'disability') AND date > '1/1/2014'	facts, holding	Work injury, labor, payment, disability, company, work injury insurance, arbitration, subsidy, salary, disposable [income]

Table 2. Topic refinement under increasingly restrictive predicate conditioning.

4. Toward Knowledge-based Retrieval and Discovery

The query and discovery infrastructure of our legal knowledge-base is polystore called AWESOME [7] which is a data management system developed over multiple data management systems including Neo4J, AsterixDB (for JSON), PostgreSQL (for relational tables) and Apache Solr (for text indexing) together with SciGraph. A detailed description of the system is beyond the scope of this paper. Here we show three examples of how the implemented knowledge-base facilitates the tasks outlined in Section 1.

Retrieval. For the retrieval query, "Find the most common government entities that are defendants in administrative litigation cases," can be executed as follows: (a) from the ontology, find all cases that are type "administrative litigation cases", (b) extract all defendants from these cases, (c) remove the particular location of government from the party's name (d) return most frequent entities that are defendants, ordered by their frequency of occurrence. The results, shown in Table 3, provide a summary of the top 10 government bodies that are the target of contention in the Chinese legal system. In particular, levels of government dealing with land, family planning, benefits, and public security are most likely to appear as defendants in these cases.

1.	People's Government	2.	Public Security Bureau
3.	Human Resources and Social Security Bureau	4.	Land and Resources Bureau
5.	Housing and Urban Construction Bureau	6.	Real Estate Authority
7.	Population and Family Planning Commission	8.	Urban and Rural Planning Bureau
9.	Administration for Industry and Commerce	10.	Real Estate Authority

Table 3. Most common government entities that are defendants of administrative litigation cases, Henan.

Discovery. Our example query, "Discover the issues of dispute for divorce-related cases where the plaintiff is female" can be interpreted as follows: (a) from the ontology, find all subclasses of case types with the term "divorce" in them (ontology fragment), (b) based on the ontology IDs of these concepts, identify the JDDs that have been marked with these IDs (mapping fragment) , (c) filter those JDDs from (b) where the value of the party with `role: plaintiff` has `gender:female` (semistructured fragment), (d) with the facts and holdings sections from cases in (c), run the topic model with an increasing number of topics, and in case store the dominant topics. A topic is called "common" if the number documents supporting it is high, and the same topic, occurs across multiple topic counts. Rarer topics are discovered as the number of topics is higher, yielding finer

1. **Child Support**	"Maintenance", "care", "child", "child development", "daughter's marriage, "life", "paid", "daughter grow up healthy"
2. **Domestic Violence**	"Neck treatment", "beat", "pinch", "drinking", "relapse", "perforation of ear and eardrum", "threatening"
3. **Division of Property**	"Washing machine", "sofa", " Cabinet", "Haier color TV", "Dresser", "coffee table", "wall units", "water dispenser"
4. **Inadequacy of Alimony**	"Income", "education", "living expenses", "custody", "unreasonable demands", "visitation rights", "born out of wedlock", "usufruct"
5. **Reconciliation**	"Tolerant", "shortcoming", "mutual trust", "communication", "harmonious", "mutual understanding", "harmonious", "exchange"

Table 4. Common (1-4) and less common (5) issues for divorce cases where the plaintiff is female.

Topic	Prop. Individual Cases	Prop. Collective Cases
Withdrawals: Withdrawn, granted, withdrawn, charged, process, examined, halved, voluntarily, should	0.20	0.18
Public Security: Penalties, transcripts, decisions, inquiries, decisions, public security, law and order, detention, management, beating	0.04	0.01
Forest Rights: contract, forest warrants, trees, awarded, contract, Li Baowei, civil, signed, woodland, publicity	0.01	0.10
Land Use Rights: land, homestead, use, use certificate, issue, dispute, use rights, collective, land, area	0.10	0.15

Table 5. Topics most associated with individual and collective plaintiffs in administrative litigation cases.

topics. In Table 4 the first 4 topics are common, while the last topic appears only when the topic count > 35, and is supported by 343 JDDs.

Legal Pattern Analysis. Last, we show how the knowledge-based representation can be used to uncover patterns in legal cases on a corpus level. To do this, we turn to our example query "Calculate the major differences between cases where plaintiffs file individually versus collectively in administrative litigation cases." To do this we (a) select administrative litigation cases (b) retrieve the parties in all selected cases (c) distinguish between those with only one plaintiff from collective parties (d) run a topic model that estimates the relationship between the topics and the plaintiff type.

Table 5 shows the topics most associated with collective plaintiffs and those most associated with individual plaintiffs. Interestingly, one of the topics more associated with individual cases than collective cases is that of case withdrawal, thought to be a shortcoming of the Chinese legal system [19]. Land cases are most likely to be filed collectively, often overy land use rights or forest rights.

5. Conclusion

We have developed an initial approach for constructing a legal knowledge base that facilitates corpus-level analyses. By combining ontologies, semistructured representations of legal sentences, and unsupervised estimation of topics on remaining unstructured data, we allow for flexible analyses that retrieve, discover, and estimate patterns at the corpus level in Chinese legal documents. Our future work includes automatic assignment

of ontological classes to JDDs using the laws applied to a case. We hope our approach provides a framework for knowledge representation that facilitates our understanding of legal systems as a whole.

References

[1] M. Alberti, A. Gomes, R. Gonçalves, J. Leite, and M. Slota. Normative systems represented as hybrid knowledge bases. *Computational Logic in Multi-Agent Systems*, pages 330–346, 2011.

[2] M. Araszkiewicz, A. Łopatkiewicz, A. Zienkiewicz, and T. Zurek. Representation of an actual divorce dispute in the parenting plan support system. In *Proc. of the 15th Int. Conf. on AI and Law*, pages 166–170. ACM, 2015.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. of Machine Learning Research*, 3(Jan):993–1022, 2003.

[4] L. K. Branting. Data-centric and logic-based models for automated legal problem solving. *Artificial Intelligence and Law*, 25(1):5–27, 2017.

[5] A. Cernian, D. Carstoiu, O. Vasilescu, and A. Olteanu. Ontolaw-ontology based legal management and information retrieval expert system. *J. of Control Engg and Applied Informatics*, 15(4):77–85, 2013.

[6] X. Chen. *Social protest and contentious authoritarianism in China*. Cambridge University Press, 2012.

[7] S. Dasgupta, K. Coakley, and A. Gupta. Analytics-driven data ingestion and derivation in the awesome polystore. In *IEEE Int. Conf. on Big Data (Big Data)*, pages 2555–2564. IEEE, 2016.

[8] M. El Ghosh, H. Naja, H. Abdulrab, and M. Khalil. Towards a legal rule-based system grounded on the integration of criminal domain ontology and rules. *Procedia Computer Science*, 112:632–642, 2017.

[9] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Pearson, 2015.

[10] L. H. Fincher. *Leftover women: The resurgence of gender inequality in China*. Zed Books Ltd., 2016.

[11] F. Galgani, P. Compton, and A. Hoffmann. Lexa: Building knowledge bases for automatic legal citation classification. *Expert Systems with Applications*, 42(17):6391–6407, 2015.

[12] T. F. Gordon. A use case analysis of legal knowledge-based systems. In *JURIX*, 2003.

[13] J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297, 2013.

[14] X. He and K. Ng. Pragmatic discourse and gender inequality in china. *Law & Society Review*, 47(2):279–310, 2013.

[15] R. Kumar and K. Raghuveer. Legal document summarization using latent dirichlet allocation. *Int. J. of Computer Science and Telecommunications*, 3:114–117, 2012.

[16] K. Li. "what he did was lawful": Divorce litigation and gender inequality in china. *Law & Policy*, 37(3):153–179, 2015.

[17] B. L. Liebman, M. Roberts, R. E. Stern, and A. Z. Wang. Mass digitization of chinese court decisions: How to use text as data in the field of chinese law. *Social Science Research Network Collection*, June 2017.

[18] M. A. Livermore, A. Riddell, and D. Rockmore. Agenda formation and the us supreme court: A topic model approach. 2016.

[19] N. Mahboubi. Suing the government in china. *Democratization in China, Korea, and Southeast Asia. Londres: Routledge*, pages 141–155, 2014.

[20] H. Nguyen and D. J. Litman. Extracting argument and domain words for identifying argument components in texts. In *Argument Mining at NAACL-Human Language Technology*, pages 22–28, 2015.

[21] Y. Panagis, M. L. Christensen, and U. Sadl. On top of topics: Leveraging topic modeling to study the dynamic case-law of international courts. In *JURIX*, pages 161–166, 2016.

[22] M. E. Roberts, B. M. Stewart, and E. M. Airoldi. A model of text for experimentation in the social sciences. *J. of the Amer. Stat. Assoc.*, 111(515):988–1003, 2016.

[23] A. Stranieri and J. Zeleznikow. The evaluation of legal knowledge based systems. In *Proc. of the 7th Int. Conf. on AI and Law*, pages 18–24, 1999.

[24] R. Winkels et al. The openlaws project: Big open legal data. In *Proc. of the 18th Int, Legal Informatics Symposium IRIS 2015*, pages 189–196, 2015.

[25] T. Zurek. Conflicts in legal knowledge base. *Foundations of Computing and Decision Sciences*, 37(2):129–145, 2012.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-145

Automated Detection of Unfair Clauses in Online Consumer Contracts

Marco LIPPI [a,1], Przemyslaw PALKA [b], Giuseppe CONTISSA [c], Francesca LAGIOIA [b], Hans-Wolfgang MICKLITZ [b], Yannis PANAGIS [d], Giovanni SARTOR [c], and Paolo TORRONI [e]

[a] *DISMI – Università di Modena e Reggio Emilia, Italy*
[b] *Law Department, European University Institute, Florence, Italy*
[c] *CIRSFID, Alma Mater – Università di Bologna, Italy*
[d] *iCourts, University of Copenhagen, Denmark*
[e] *DISI, Alma Mater – Università di Bologna, Italy*

Abstract. Consumer contracts too often present clauses that are potentially unfair to the subscriber. We present an experimental study where machine learning is employed to automatically detect such potentially unfair clauses in online contracts. Results show that the proposed system could provide a valuable tool for lawyers and consumers alike.

Keywords. Unfair terms detection, Consumer contract, Machine learning

1. Introduction

A PhD student from Poland plans to move to Italy. She will have to open a bank account, rent a flat, get a local phone number, etc. She will have to sign many lengthy contracts. Most of them will be only written in Italian. Can she simply focus on the costs and features of services described in the contracts? Or will she have to worry about possible 'legal traps' as well?

It is a fact that consumers rarely read the contracts they are required to accept [19], and even if they do, they have no means to influence their content. This created a need for limitations on contractual freedom [13], not only to protect consumer interests, but also to enhance the consumers' trust in transnational transactions and improve the common market [18]. The same considerations apply to online platforms, a necessary component of Junker Commission's Digital Single Market initiative.[2] Because consumers cannot realistically be expected to read and fully understand all the contracts they sign, European consumer law aims to prevent businesses from using so-called 'unfair contractual terms' in the contracts they unilaterally draft and require consumers to accept [20]. Law regarding such terms applies also to the Terms of Service (ToS) of online platforms [12]. Unfor-

[1] Corresponding Author: marco.lippi@unimore.it.
[2] Brussels, 6.5.2015COM(2015) 192 final. Communication: A Digital Single Market Strategy for Europe.

tunately, it turns out that owners of these platforms, such as Google, Facebook and Twitter, do use in their ToS unfair contractual clauses, in spite of the European law, and regardless of consumer protection organizations, which have the competence, but not the resources, to fight against such unlawful practices.

We propose to address this problem by partially automating the detection of (potentially) unfair clauses using machine learning. This paper follows and combines results of our earlier work. That includes an analysis of the legal issues involved in the automation of enforcement of consumer law regarding unfair contractual clauses, and have developed a software that detects unfair clauses, based on manually created rules encoding recurring textual structures, which gave promising results [15]. However, such an approach has a drawback, in that it is labor-intensive and struggles to cope with the diversity and rapid evolution of the language of ToS. In other recent work we trained a machine learning classifier on a corpus annotated by domain experts, and successfully used it to extract claims from legal documents [11]. Here we build on the work done so far by applying machine learning methods to the detection of unfair contract clauses.

We have structured this paper as follows. In Section 2 we introduce the legal problem. Section 3 describes the corpus and the document annotation procedure. Section 4 explains the machine learning methodology employed in the system, whereas Section 5 presents experimental results. Section 6 discusses related work and concludes with a look to future research.

2. Problem Description

In this section we briefly introduce the European consumer law on unfair contractual terms (clauses). We explain what an unfair contractual term is, present the legal mechanisms created to prevent business from employing them, and describe how our project will contribute to these mechanisms.

According to art. 3 of the Directive 93/13 on Unfair Terms in Consumer Contracts, a contractual term is unfair if: 1) it has not been individually negotiated; and 2) contrary to the requirement of good faith, it causes a significant imbalance in the parties rights and obligations, to the detriment of the consumer. This general definition is specified in the Annex to the Directive, containing an indicative and non-exhaustive list of the terms which may be regarded as unfair, as well in by more than 50 judgments of the Court of Justice of the EU [14]. Examples of unfair clauses encompass taking jurisdiction away from the consumer, limiting liability for damages on health and/or gross negligence, imposing obligatory arbitration in a country different from consumers residence etc.

Loos and Luzak [12] identified five categories of *potentially unfair* clauses: 1) establishing jurisdiction for disputes in a country different than consumers residence; 2) choice of a foreign law governing the contract; 3) limitation of liability; 4) the provider's right to unilaterally terminate the contract/access to the service; and 5) the provider's right to unilaterally modify the contract/the service itself. Our research identified three additional categories: 6) requiring a consumer to undertake arbitration before the court proceedings can commence; 7) the provider retaining the right to unilaterally remove consumer content from the service; 8)

having a consumer accept the agreement simply by using the service, not only without reading it, but even without having to click on "I agree/I accept."

The 93/13 Directive creates two mechanisms to prevent the use of unfair contractual terms: *individual* and *abstract* control of fairness. The former takes place when a consumer goes to court: if a court finds that a clauses is unfair (which it can do on its own motion), it will consider that the clause is not binding on the consumer (art. 6). However, most consumers do not take their disputes to courts. That is why abstract fairness control has been created. In each EU Member State, consumer protection organizations have the competence to initiate legal proceedings aiming to obtain the declaration that clauses in consumer contracts are unfair, through judicial or in administrative proceedings. The national implementations of abstract control may differ—public authorities or civil society organizations may be involved; there may or may not be fines for using unfair contractual terms; etc. [21]—but what is common to all member states is that if a business uses unfair terms in their contracts, in principle there is always someone competent to make them stop.

Unfortunately, the legal mechanism for enforcing the prohibition of unfair contract terms have been unable to effectively counter this practice so far. As reported by some literature [12], and as our own research indicates [15], unfair contractual terms are, as of today, widely used in ToS of online platforms.

In our previous research [15] we developed a theoretical model of tasks that human lawyers currently need to carry out before legal proceedings concerning the abstract control of fairness of clauses can begin. Those include: 1) finding and choosing the documents; 2) mining the documents for potentially unfair clauses; 3) conducting the actual legal assessment of fairness; 4) drafting the case files and beginning the proceedings. Our project aims to automate the second step, enabling a senior lawyer to focus only on clauses that are found by a machine learning classifier to be potentially unfair, thus saving significant time and labor. Our classifiers will look not only for clearly unfair clauses but also for *potentially* unfair ones. The focus on potentially unfair clauses is due to two main reasons.

First, we may be uncertain on whether a certain type of clause falls under the abstract legislative definition of an "unfair contractual term". One can only have legal certainty that a certain type of clause is unfair if a competent institution, such as the European Court of Justice, has decided so. That is the case for certain kinds of clauses, such as a jurisdiction clause indicating a country different from the consumer's residence, or limitation of liability for gross negligence [15]. In other cases the unfairness of a clause, has to be argued for, showing that it creates an unacceptable imbalance in the parties' rights and obligations. A consumer protection body might want to take the case to a court in order to authoritatively establish the unfairness of that clause, but a legal argument for that needs to be created, and the clause may eventually turn out to be judged fair.

Second, we may remain uncertain on the unfairness of a particular clause detected by the classifier, since its unfairness may depend not only on its textual content, but also on the context in which the clause is to be applied. For instance, a mutual right to unilaterally terminate the contract might be fair in some cases, and unfair in others, for example if unilateral termination would entail losing some digital content (purchased apps, email address, etc.) on the side of the consumer.

3. Corpus Annotation

In order to train machine learning classifiers we produced a corpus consisting of 20 relevant on-line consumer contracts, i.e. the ToS of the following on-line platforms: 9gag.com, Academia.edu, Amazon, eBay, Dropbox, Facebook, Google, Linden Lab, Microsoft, Netflix, Rovio, Snapchat, Spotify, Supercell, Twitter, Vimeo, World of Warcraft, Yahoo, YouTube and Zynga. When more than one version of the same contract was available, we selected the most recent version available on-line for the European customers. The corpus contains overall 5,103 sentences, 333 of which we marked as expressing (potentially) unfair clauses. If a clause span included multiple sentences, we decided to tag all such sentences. We used XML as a mark-up language.

An initial analysis of our corpus enabled us to identify 8 different types of clause, for which we defined 8 corresponding XML tags: jurisdiction (`<j>`), choice of law (`<law>`), limitation of liability (`<ltd>`), unilateral change (`<ch>`), unilateral termination (`<ter>`), arbitration (`<a>`), contract by using (`<use>`), and content removal (`<cr>`). We assumed that for each type of clause we could distinguish three classes: (a) clearly fair, (b) potentially unfair, and (c) clearly unfair. In order to mark the different degrees of (un)fairness we appended a numeric value to each XML tag, with 1, 2, and 3, meaning clearly fair, potentially unfair, and clearly unfair respectively. For instance, the tag `<j3>` indicates that the tagged clause is classified as a clearly unfair jurisdiction clause.

A **jurisdiction** clause specifies what courts will adjudicate the disputes arising from the contract. If a jurisdiction clause gave consumers the right to bring disputes in their place of residence, the clause was marked as clearly fair, whereas it was marked as clearly unfair if it stated that any judicial proceeding takes a residence away (i.e. in a different city, different country). As an example consider the following clauses taken from the Dropbox ToS:

```
<j3>You and Dropbox agree that any judicial proceeding to resolve claims
relating to these Terms or the Services will be brought in the federal or state
courts of San Francisco County, California [...]</j3>

<j1>If you reside in a country (for example, European Union member states) with
laws that give consumers the right to bring disputes in their local courts,
this paragraph doesn't affect those requirements.</j1>
```

The second clause introduces an exception to the general rule stated in the first clause, so the first one was marked as clearly unfair and the second as clearly fair.

A **choice of law** clause specifies which law will govern the relations arising from the agreement, and according to which law a potential dispute will be adjudicated. If the applicable law was determined based on the consumer's country of residence, the clause was marked as clearly fair. In any other case the choice of law clause was considered to be potentially unfair. The following example is taken from the Facebook ToS:

```
<law2>The laws of the State of California will govern this Statement, as well
as any claim that might arise between you and us, without regard to conflict of
law provisions</law2>
```

A **limitation of liability** clause specifies the amount and types of damages that the service provider will be obligated to provide to consumers under terms and conditions stipulated in the service agreement. Clauses that did not exclude or limit the liability were marked as clearly fair. Potential unfairness was attributed to clauses that reduced, limited, or excluded the liability of the service provider for damages (such as any harm to the computer system because of malware or loss of data), and for the suspension, modification, discontinuance or lack of the availability of the service. This classification was also applied to clauses as well as those containing blanket phrases like "to the fullest extent permissible by law". Clauses that reduced, limited, or excluded the liability of the service provider for physical injuries, intentional damages as well as in case of gross negligence, were marked as clearly unfair.

A **unilateral change** clause in favour of the provider specifies the conditions under which the service provider can amend and modify the ToS. Such clauses were consistently marked as potentially unfair.

A **unilateral termination** in favour of the provider details the circumstances under which the provider can suspend and/or terminate the service and/or the contract. We marked such clauses as follows: potentially unfair if the suspension or termination was allowed only under specific reasons and conditions; clearly unfair if they empowered the service provider to suspend or terminate the service at any time for any or no reasons and/or without notice. That was the case in the Academia terms of use:

```
<ter3>Academia.edu reserves the right, at its sole discretion, to discontinue
or terminate the Site and Services and to terminate these Terms, at any time
and without prior notice.</ter3>
```

A **contract by using** clause states that the consumer is bounded by the terms of use of a specific service, simply by using the service. We consistently marked such clauses as potentially unfair.

A **content removal** clause specifies the conditions under which the service provider may remove the user's content. We marked the clause as follows: potentially unfair if the clause specified reasons and conditions for such a removal; clearly unfair if it stated that the provider may remove content in his full discretion, and/or at any time for any or no reasons and/or without notice nor possibility to retrieve the content.

Finally, an **arbitration** clause requires the parties to resolve their disputes through an arbitration process, before the case can go to court. Tt is thus considered as a kind of forum selection clause. Such a clause may or may not specify that arbitration occur within a specific jurisdiction. We marked such a clause as follows: clearly fair if it defined the arbitration as fully optional; clearly unfair if it stated that the arbitration (1) takes place in a state other then the state of consumer's residence and/or (2) it is not based on law but on arbiter's direction; potentially unfair in all other cases.

4. Machine Learning Methodology

From a machine learning point of view, the problem of detecting unfair clauses within a contract can be seen as a sentence classification task. Given a sentence belonging to a contract, the goal is to classify it as *positive* (if the sentence expresses a clearly or potentially unfair clause) or *negative* (otherwise). In order to train a machine learning system able to distinguish positive from negative sentences, a supervised learning algorithm is typically employed. This framework assumes the availability of a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ made up of N pairs (x_i, y_i) where x_i is the representation of a sentence, and y_i is its corresponding label (also named class or target), which is the category to be predicted. In this case, we consider only two sentence categories, namely positive (clearly or potentially unfair) and negative (clearly fair), thus we deal with a binary classification task.

There are many approaches for the classification of sentences. In this paper we consider and compare two of them. The first one, known as *bag-of-words* (BoW), consists in representing a sentence as a vector of features that is as large as the dimension of the vocabulary of words within the dataset. Each feature is either zero (if the corresponding word does not appear in the sentence) or different from zero (if it does). The non-zero value in the feature vector associated to each word is the so-called TF-IDF score, that is the number of times the word appears in the sentence (Term Frequency, TF) multiplied by a term that amplifies the contribution of rare words (Inverse Document Frequency, IDF) [22]. A sentence representation, such as its BoW, can be fed to different types of machine learning classifiers. In this work we employ support vector machines (SVMs), as they are widely used in text classification [6]. Extensions of the BoW approach consider so-called n-grams, i.e. features extracted from the sentence by taking into account the frequencies of consecutive word combinations, and grammatical information such as part-of-speech tags, i.e., word categories such as nouns, verbs, etc. [7]. The BoW approach is thus built to leverage the *lexical* information within a sentence, and in particular the presence of keywords and phrases that are highly discriminative for the detection of unfair clauses.

The second approach we consider in our study is that of *tree kernels* [17] (TK). This approach takes into account the similarity between the *structure* of sentences and has been shown to offer state-of-the-art performance in related classification tasks, such as those typical of argumentation mining [10], for example claim detection [8]. The structure of a sentence is naturally encoded by its *constituency parse tree*, which describes the syntactic and grammatical characteristics of a sentence. A TK consists of a *similarity measure* between two trees, by taking into account the number of common substructures or *fragments*. Different definitions of fragments induce different TK functions. In our study we use the SubSet Tree Kernel (SSTK) [4] which counts as fragments those subtrees of the constituency parse tree terminating either at leaves or the level of non-terminal symbols. SSTK have been shown to outperform other TK functions in several argumentation mining sub-tasks [9].

5. Experimental Results

We performed experiments on the dataset described in Section 3, following a standard *leave-one-document-out* (LOO) procedure, whereby each document in the corpus is used, in turn, as test set for our classifier, while the remaining documents constitute the training set. In this way, we obtain predictions for each document in the dataset, and we measure the performance on each contract separately, thus evaluating the generalization capabilities of the system. In particular, we compute precision, recall and F_1 for each contract, and we finally compute the average for each of these three metrics (this is called macro-average [22]). Precision (P) is defined as the fraction of examples predicted as positives, which are actually labeled as positive. Recall (R) is the fraction of positive examples that are correctly detected. F_1 is finally the harmonic mean between precision and recall ($F_1 = \frac{2PR}{P+R}$).

As customary in studies of this kind, the above performance measures are compared with baselines that give an indication of the difficulty of the problem at hand. We aim to compare three systems:

1. a single SVM exploiting BoW (unigrams and bigrams), considering as the positive class the union of all tagged sentences;
2. a combination of eight SVMs exploiting the same features as above, but each considering as the positive class only one specific tag; a sentence is then predicted as unfair if at least one of the SVMs predicts it as such;
3. a kernel machine exploiting TK, considering as the positive class the union of all tagged sentences.

We adopt two standard baselines: a *random* baseline, which predicts unfair clauses at random,[3] and an *always positive* baseline, which predicts every sentence as unfair. If any of these baselines provided a result with acceptable accuracy, that would mean that the classification task has a trivial solution.

Table 1 shows the results achieved by each of these variants. We notice that the precision of baseline classifiers is below 8%, and that the precision of either BoW and TK is above 57%. Moreover, we notice how the single-model SVM performs best, outperforming both Tree Kernels, which exploit the same setting for the definition of positive class, and the combined-model SVM, which separately trains a different model for each category (tag) of unfair clauses.

These figures tell us something about the nature of the task. First, the better performance of the single model with respect to the combined model implies that knowing unfair clauses of different categories is useful to correctly predict the unfair clauses of a specific category. This is particularly important for corpora where few tagged examples exist for a certain category, but it is also interesting from a computational linguistic and legal point of view, since it seems to suggest the existence of *a common lexicon for unfair clauses*, which spans across several tag categories. Second, the worse performance associated with TK suggests that the syntactic structure of the sentence is probably not very indicative of the presence of an unfair clause—or, at least, that it is less informative than the lexical information captured by n-grams. This makes the task of detecting unfair

[3]Sampling takes into account the class distribution in the training set.

Table 1. Results on leave-one-document-out procedure.

Method	P	R	F_1
SVM – Single Model	0.620	0.715	0.648
SVM – Combined Model	0.576	0.621	0.582
Tree Kernels	0.571	0.665	0.603
Random Baseline	0.071	0.071	0.071
Always Positive Baseline	0.075	1.000	0.138

Table 2. Recall of abusive clauses for each tag category for the single and combined SVM models, micro-averaged on the whole dataset.

Tag	Single	Combined
Arbitration	0.531	0.344
Unilateral change	0.809	0.723
Content removal	0.677	0.645
Jurisdiction	0.826	0.826
Choice of law	1.000	0.778
Limitation of liability	0.614	0.602
Unilateral termination	0.780	0.744
Contract by using	0.579	0.342

clauses different from other text retrieval problems in the legal domain, such as, for example, the detection of claims and arguments [10].

In Table 2 we also report the recall of the single- and combined-model SVM for each separate tag category, micro-averaged on the whole dataset. The results show that all the categories benefit from the knowledge of unfairness given by the other categories: this is particularly significant for the "Arbitration" and "Contract by using" categories, which still remain the hardest to detect.

Interestingly, preliminary experimental results provided some feedback to the tagging: a number of apparent false positives where due to mistakes in tagging; they concerned unfair clauses that had escaped the analysts, due to the length and complexity of the ToS.

6. Conclusions

The use of machine learning and natural language processing techniques in the analysis and classification of legal documents is gaining a growing interest. Moens et al. [16] proposed a pipeline of steps for the extraction of arguments from legal documents, exploiting supervised classifiers and context-free grammars, whereas Biagioli et al. [3] proposed to employ multi-class SVM for the identification of significant text portions in normative texts. Recent approaches have focussed on the detection of claims [11] and of cited facts and principles in legal judgments [23], as well as on the prediction of judicial decisions [1]. A case study regarding the construction of legal arguments in the legal determinations of vaccine/injury compensation compliance using natural language tools was given in [2]. It is worth remarking that, in most of these works, classic lexical features such as BoW still represent a crucial ingredient of automated systems. Finally, privacy policies rep-

resent another strictly related application where machine learning approaches have proved effective (e.g., see [5] and references therein).

This paper presented a first experimental study that used machine learning to address the automated detection of potentially unfair clauses in online contracts. Our results seem encouraging: using a small training set we could automatically detect most unfair clauses, and with acceptable precision. Given that most unfair clauses are currently hidden within long and hardly readable ToS, the recall and precision offered by our approach may be already sufficient for practical applications.

Interesting and to some extent unexpected outcomes included the comparatively better performance of the BoW approach, and the fact that the automated detection method we developed was able to highlight a number of unfair clauses that human analysts had failed to identify in the first place.

This study was motivated by a long-term goal such as the pursuit of effective consumer protection by way of tools that support consumers and their organizations in detecting unfair contractual clauses. That is also the objective of a research and development project (CLAUDETTE) that has recently kicked off at the European University Institute. Looking to the future, we plan to carry out further analyses that enable us to determine what machine learning methods should be implemented in such future tools. Accordingly, we plan to conduct a qualitative analysis of the errors performed by our system, in order to identify weaknesses and improve performance. We are also working on the construction of a larger corpus, with the intention of improving training as well as providing a suitable dataset for testing other machine learning algorithms, such as deep networks, which have proven effective in several other natural language processing tasks. Finally, we are studying ways to exploit contextual information, since it was pointed out that a clause might be fair in a context but not in others.

References

[1] N. Aletras, D. Tsarapatsanis, D. Preoiuc-Pietro, and V. Lampos. Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. *PeerJ Computer Science*, 2:e93, Oct. 2016.

[2] K. D. Ashley and V. R. Walker. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 176–180. ACM, 2013.

[3] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. Automatic semantics extraction in law documents. In *Proceedings of ICAIL*, pages 133–140. ACM, 2005.

[4] M. Collins and N. Duffy. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 263–270. ACL, 2002.

[5] B. Fabian, T. Ermakova, and T. Lentz. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*, pages 18–25. ACM, 2017.

[6] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *ECML 98*, pages 137–142, 1998.
[7] E. Leopold and J. Kindermann. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3): 423–444, 2002.
[8] M. Lippi and P. Torroni. Context-independent claim detection for argument mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 185–191, 2015.
[9] M. Lippi and P. Torroni. Margot. *Expert Syst. Appl.*, 65(C):292–303, Dec. 2016. ISSN 0957-4174. doi: 10.1016/j.eswa.2016.08.050.
[10] M. Lippi and P. Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25, Mar. 2016.
[11] M. Lippi, F. Lagioia, G. Contissa, G. Sartor, and P. Torroni. Claim detection in judgments of the EU Court of Justice. In *VI Int. Workshop on Artificial Intelligence and the Complexity of Legal Systems (AICOL)*, 2015.
[12] M. Loos and J. Luzak. Wanted: a bigger stick. on unfair terms in consumer contracts with online service providers. *Journal of Consumer Policy*, 39(1): 63–90, 2016.
[13] H.-W. Micklitz. On the intellectual history of freedom of contract and regulation. *Penn State Journal of Law & International A airs*, 4(1), Dec. 2015.
[14] H.-W. Micklitz and N. Reich. The court and sleeping beauty: The revival of the unfair contract terms directive (uctd). *Common Market Law Review*, 51 (3):771–808, 2014.
[15] H.-W. Micklitz, P. Pałka, and Y. Panagis. The empire strikes back: Digital control of unfair terms of online services. *Journal of Consumer Policy*, pages 1–22, 2017.
[16] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM, 2007.
[17] A. Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In J. Frnkranz, T. Scheffer, and M. Spiliopoulou, editors, *ECML 2006*. Springer Berlin Heidelberg, 2006.
[18] P. Nebbia. *Unfair contract terms in European law: a study in comparative and EC law*. Bloomsbury Publishing, 2007.
[19] J. A. Obar and A. Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. 2016.
[20] N. Reich, H.-W. Micklitz, P. Rott, and K. Tonner. *European consumer law*. Intersentia, 2014.
[21] H. Schulte-Nölke, C. Twigg-Flesner, and M. Ebers. *EC consumer law compendium: the consumer acquis and its transposition in the member states*. Walter de Gruyter, 2008.
[22] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.
[23] O. Shulayeva, A. Siddharthan, and A. Wyner. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, 25(1):107–126, Mar 2017.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-155

A Deep Learning Approach to Contract Element Extraction

Ilias Chalkidis [a,b], Ion Androutsopoulos [a]
[a] *Athens University of Economics and Business, Greece*
[b] *Cognitiv+ Ltd., London, UK*

Abstract. We explore how deep learning methods can be used for contract element extraction. We show that a BILSTM operating on word, POS tag, and token-shape embeddings outperforms the linear sliding-window classifiers of our previous work, without any manually written rules. Further improvements are observed by stacking an additional LSTM on top of the BILSTM, or by adding a CRF layer on top of the BILSTM. The stacked BILSTM-LSTM misclassifies fewer tokens, but the BILSTM-CRF combination performs better when methods are evaluated for their ability to extract entire, possibly multi-token contract elements.

Keywords. Natural language processing, deep learning, legal text analytics.

1. Introduction

Law firms, companies, government agencies etc. need to monitor contracts for a wide range of tasks [1]. For example, law firms need to inform their clients when contracts are about to expire or when they are affected by legislation changes. Contractors need to keep track of agreed payments and deliverables. Law enforcement agencies may need to focus on contracts involving particular parties and large payments. Many of these tasks can be automated by extracting particular contract elements (e.g., termination dates, legislation references, contracting parties, agreed payments). Contract element extraction, however, is currently performed mostly manually, which is tedious and costly.

We recently released a benchmark dataset of approximately 3,500 English contracts, annotated with 11 types of contract elements, the largest publicly available dataset for contract element extraction.[1] Using that dataset, in previous work [2] we experimented with Logistic Regression [3] and linear Support Vector Machines (SVMs) [4], both operating on fixed-size sliding windows of tokens, represented using hand-crafted features, pre-trained word embeddings [5,6], and/or pre-trained part-of-speech (POS) tag embeddings. We also experimented with manually written rules that replaced the machine learning classifiers or post-processed their decisions. In this paper, we experiment with deep learning methods [7,8] on the same dataset. We show that a bidirectional LSTM (BILSTM) [9,10,11] operating on word, POS tag, and token-shape embeddings outperforms the best methods of our previous work, in most cases, without using any manually written rules. Further improvements are observed by stacking an additional LSTM on top of the BIL-

[1] The dataset is available from `http://nlp.cs.aueb.gr/publications.html`.

STM [12,13] or by adding a Conditional Random Field (CRF) layer [14] on top of the BILSTM [15,16,17]. The stacked BILSTM-LSTM misclassifies fewer tokens, but the BILSTM-CRF combination performs better when methods are evaluated for their ability to extract entire, possibly multi-token contract elements.

2. Contract Element Extraction Methods

The dataset of our previous work specifies particular *extraction zones* for each contract element type [2]. For example, contracting parties are to be extracted from the cover page and preamble of each contract, whereas legislation references are to be extracted from zones starting up to 20 tokens before and ending up to 20 tokens after each occurrence of words like "act" or "treaty" in the main text. The extraction zones are explicitly marked in each training and test contract of the dataset, and can be easily produced in practice (e.g., using regular expressions). We use the same extraction zones in this paper.

We also use the pre-trained 200-dimensional word embeddings and 25-dimensional POS tag embeddings that accompany the dataset [2], which were obtained by applying WORD2VEC (skip-gram model) [18] to approximately 750,000 unlabeled and 50,000 POS-tagged English contracts, respectively. We also use 5-dimensional *token shape embeddings* that represent the following seven possible shapes of tokens: token consisting of alphabetic upper case characters, possibly including periods and hyphens (e.g., 'AGREEMENT', 'U.S.', 'CO-OPERATION'); token consisting of alphabetic lower case characters, possibly including periods and hyphens (e.g., 'registered', 'etc.', 'third-party'); token with at least two characters, consisting of an alphabetic upper case first character, followed by alphabetic lower-case characters, possibly including periods and hyphens (e.g., 'Limited', 'Inc.', 'E-commerce'); token consisting of digits, possibly including periods and commas (e.g., '2009', '12,000', '1.1'); line break; any other token containing only non-alphanumeric characters (e.g., '$', '##'); any other token (e.g., '3rd', 'strangeTek', 'EC2'). The token shape embeddings were obtained by applying WORD2VEC (again, skip-gram model, same other settings) to approx. 2,000 contracts of the unlabeled dataset of our previous work [2], after replacing the tokens by pseudo-tokens (e.g., 'allupper', 'alllower') reflecting the corresponding token shape.[2]

As already noted, in our previous work we experimented with linear classifiers (Logistic Regression and linear SVMs) operating on fixed-size sliding windows of tokens. We also experimented with manually written rules that replaced the linear classifiers or post-processed their decisions. Those methods are described in detail in our previous work [2]. Below we describe the new, LSTM-based methods we experimented with. We do not consider manually written rules in any detail, because in most cases the LSTM-based methods outperform our previous methods (with or without rules) without employing any rules. As in our previous work, for each LSTM-based method we build a separate extractor for each contract element type (e.g., contracting parties), 11 extractors in total per method, which allows us to compare directly against our previous results.[3]

[2] To by-pass privacy issues, in the dataset each token is replaced by a unique integer identifier, but hand-crafted features, word, and POS tag embeddings are still provided for each token [2]. The new token shape embeddings will also be made publicly available. We use no hand-crafted features in this paper.

[3] The LSTM-based methods were implemented using KERAS (`https://keras.io/`) with a TENSORFLOW backend (`https://www.tensorflow.org/`).

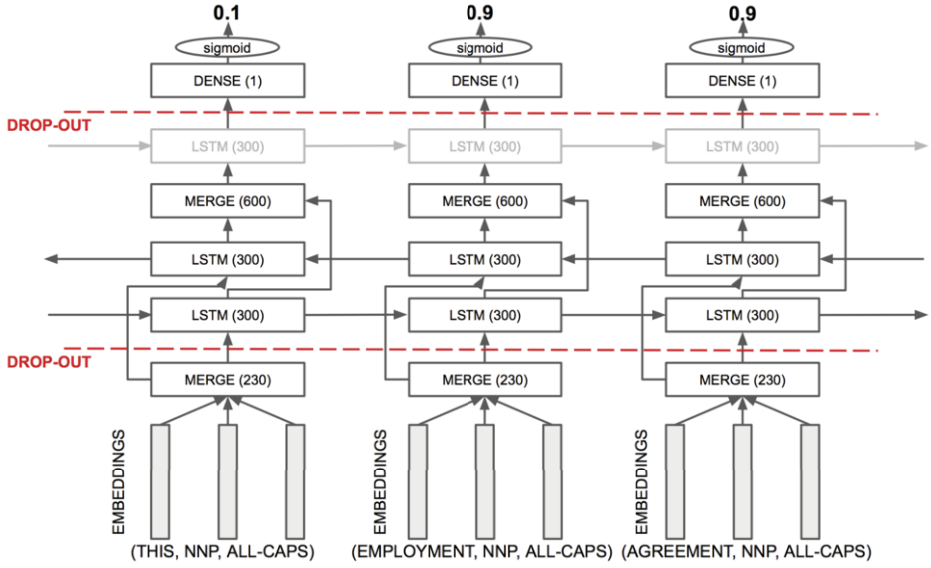


Figure 1. BILSTM-(LSTM)-LR extractor for a particular contract element type.

2.1. BILSTM-LR Extractors

In the first LSTM-based method, called BILSTM-LR, each extractor (Fig. 1, without the upper LSTM boxes) uses its own bidirectional LSTM (BILSTM) chain [12] to convert the concatenated word, POS tag, and token shape embeddings of each token (lower MERGE boxes) of an extraction zone to context-aware token embeddings (upper MERGE boxes). Each context-aware token embedding is then passed on to a Logistic Regression (LR) layer (DENSE boxes and sigmoid ovals) to estimate the probability that the corresponding token is positive (e.g., part of a contracting party element) with respect to the contract element type of the particular extractor.

We use 300-dimensional hidden states in both LSTM chains. Larger dimensionalities slow down our experiments, without noticeable efficacy improvements. We employ LSTM cells with input, forget, and output gates [9,10,19], with DROPOUT [20] after the merged embeddings and before the LR layer (Fig. 1). We used Glorot initialization [21], binary cross-entropy loss, and the Adam optimizer [22] to train each BILSTM-LR extractor, with early stopping examining the validation loss. The DROPOUT rate, learning rate, and batch size (possibly different per contract element type) were tuned performing a 3-fold cross-validation on 80% of the training extraction zones (of the corresponding contract element type), using one third of the 80% of the training extraction zones as a validation set in each fold. Having selected DROPOUT rate, learning rate, and batch size (per contract element type), each BILSTM-LR extractor was re-trained on the entire 80% of the training extraction zones, using the remaining 20% as a validation set. Out of vocabulary words, meaning words we had no pre-trained embeddings for, were mapped to random embeddings, as in our previous work. At test time, each token is classified as positive if the corresponding probability of the LR layer (Fig. 1) exceeds 0.5.

2.2. BILSTM-LSTM-LR Extractors

The second LSTM-based method, BILSTM-LSTM-LR, is the same as the previous one, except that it has an additional LSTM chain (upper LSTM boxes in Fig. 1) between the context-aware token embeddings (MERGE (600) boxes) of the lower BILSTM chain, and the logistic regression (LR) layer (DENSE boxes and sigmoid ovals). Stacking LSTM (or BILSTM) chains has been reported to improve efficacy in several linguistic tasks [13,23] at the expense of increased computational cost. To reduce the computational cost, it is common to make the stacked LSTM chains unidirectional, rather than bidirectional [23]. Hyper-parameter tuning and training are performed as in BILSTM-LR (Section 2.1).

2.3. BILSTM-CRF Extractors

In the third LSTM-based method, BILSTM-CRF, we replace the upper LSTM chain and the LR layer of the BILSTM-LSTM-LR extractor (upper LSTM and DENSE boxes, sigmoid ovals of Fig. 1) by a linear-chain Conditional Random Field (CRF). CRFs [14] have been widely used in sequence labeling (e.g., POS tagging, named entity recognition). They have also shown promising results on top of LSTM, BILSTM, or feed-forward neural network layers in sequence labeling [24,25,15,16,17] and parsing [26]. In our case, the CRF layer jointly selects the assignment of positive or negative labels to the entire token sequence of an extraction zone, which allows taking into account the predicted labels of neighboring tokens. For example, if both the previous and the next token of the current token are classified as parts of a legislation reference, this may be an indication that the current token is also part of the same legislation reference.

Again, we train a separate BILSTM-CRF extractor per contract element type. Training combines dynamic programming or beam search decoding with backpropagation to maximize log-likelihood [25,26].[4] Hyper-parameter tuning and training are performed as in the previous methods. We note that in tasks with richer sets of labels, as opposed to our only two labels ('positive', 'negative'), a CRF layer may be more beneficial. For example, in POS tagging [15,17] a CRF layer can learn that a determiner is usually followed by an adjective or noun, rather than a verb. One way to enrich our label set would be to use a single classifier for all the contract element types. This would be complicated, however, by the fact that contract elements of different types may have different extraction zones.

3. Experimental Results

We performed two groups of experiments, reported in turn below, where the methods of Section 2 where evaluated per token and contract element, respectively.

3.1. Evaluation per Token

In the first group of experiments, we evaluated the methods by considering their decisions *per token*. For each contract element type (e.g., contracting parties), we measured the

[4] We use the CRF layer implementation of KERAS-CONTRIB (`https://github.com/farizrahman4u/keras-contrib/blob/master/keras_contrib/layers/crf.py`), with joint conditional log-likelihood optimization and Viterbi best path prediction (decoding).

ELEMENT	SW-LR-ALL			SW-SVM-ALL			BILSTM-LR			BILSTM-LSTM-LR			BILSTM-CRF		
TYPE	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Title	0.91	0.91	0.91	0.91	0.91	0.91	0.95	0.93	0.94	0.94	**0.95**	**0.95**	**0.96**	**0.95**	**0.95**
Parties	0.92	0.85	0.89	0.92	0.87	0.89	0.97	0.92	**0.95**	0.97	**0.94**	**0.95**	**0.98**	0.92	**0.95**
Start	0.79	0.96	0.87	0.78	0.96	0.86	0.91	0.97	0.94	**0.93**	0.97	**0.95**	0.92	**0.98**	**0.95**
Effective	0.71	0.63	0.67	0.67	0.79	0.72	**0.98**	0.92	0.95	0.97	**0.96**	**0.97**	0.95	0.89	0.92
Termination	0.68	0.86	0.76	0.54	**0.95**	0.69	0.65	0.90	0.75	**0.70**	0.92	**0.79**	0.65	0.93	0.77
Period	**0.61**	0.74	**0.67**	0.55	0.83	0.66	0.44	0.82	0.57	0.47	**0.86**	0.59	0.55	0.85	0.65
Value	0.70	0.56	0.62	0.68	0.61	0.64	**0.74**	0.55	0.63	**0.74**	**0.63**	**0.68**	0.72	0.60	0.66
Gov. Law	0.92	0.96	0.94	0.91	0.97	0.94	0.98	**0.98**	0.98	**0.99**	**0.98**	**0.98**	**0.99**	0.97	**0.98**
Jurisdiction	0.86	0.77	0.81	0.82	0.82	0.82	**0.90**	**0.89**	**0.89**	**0.90**	0.88	**0.89**	**0.90**	0.88	0.88
Legisl. Refs.	0.84	0.83	0.83	0.83	0.88	0.86	0.82	**0.95**	**0.88**	0.83	0.94	**0.88**	0.82	0.94	0.87
Headings	0.71	0.92	0.80	0.71	0.92	0.80	0.98	0.97	**0.98**	**0.99**	**0.98**	**0.98**	**0.99**	0.97	**0.98**
Macro-average	0.79	0.82	0.80	0.76	0.86	0.80	0.85	0.89	0.86	**0.86**	**0.91**	**0.87**	**0.86**	0.90	**0.87**

Table 1. Precision (P), Recall (R), and F_1 score, *measured per token*. Best results per contract element type shown in bold font in gray cells.

performance of each method in terms of *precision* ($P = \frac{TP}{TP+FP}$), *recall* ($R = \frac{TP}{TP+FN}$), and F_1 score ($F_1 = \frac{2 \cdot P \cdot R}{P+R}$). Here, true positives (*TP*) are tokens correctly classified as parts of contract elements of the considered type, *false positives* (*FP*) are tokens incorrectly classified as parts of contract elements of the considered type, and *false negatives* (*FN*) are tokens incorrectly classified as not parts of contract elements of the considered type. F_1 (harmonic mean) is commonly used to combine precision and recall.

Table 1 lists the results of this group of experiments. The best results per contract element type are shown in bold font in gray cells. The *macro-averages* are the averages of the corresponding columns, indicating the overall performance of each method on all the contract element types. The best methods of our previous work in these experiments [2] are SW-LR-ALL and SW-SVM-ALL, which use hand-crafted features, word, and POS tag embeddings, with LR or SVM classifiers operating on fixed-size windows of tokens. Overall, both of these methods perform equally well (0.80 macro-averaged F_1).

The three LSTM-based methods overall perform clearly better than the linear sliding-window classifiers of our previous work. Even BILSTM-LR, the simplest of the three LSTM-based methods, exceeds the macro-averaged F_1 score of the best previous methods by 6 points (0.86 vs. 0.80). The extra LSTM layer of BILSTM-LSTM-LR improves the macro-averaged F_1 score by only 1 point (0.87). By looking at the results for individual contract element types, however, we see that BILSTM-LSTM-LR obtains top F_1 scores for all but one contract element types (the exception being contract periods), and for some element types (most notably, termination dates and contract values) it performs significantly better than BILSTM-LR (0.79 vs. 0.75, and 0.68 vs. 0.63 F_1, respectively). Although BILSTM-CRF has the same macro-averaged F_1 as BILSTM-LSTM-LR (0.87), it does not perform better than BILSTM-LSTM-LR in any contract element type, except for contract periods, where it outperforms BILSTM-LSTM-LR (0.65 vs. 0.59 F_1). The best F_1 for contract periods, however, is achieved by SW-LR-ALL (0.67); SW-SVM-ALL (0.66 F_1) also exceeds the F_1 score of BILSTM-CRF (0.65) for contract periods.

The lowest F_1 scores of all three LSTM-based methods are for contract periods, termination dates, and contract values, which are the three contract element types with the fewest training instances in the dataset [2]. The performance of the best sliding-window methods (SW-LR-ALL, SW-SVM-ALL) is close to or better than the performance of BILSTM-LR (the weakest LSTM-based method) in these three contract element types;

and both BILSTM-LSTM-LR and BILSTM-CRF show some of their biggest improvements compared to the simpler BILSTM-LR in these three types.

It seems that the LSTM-based methods perform poorly for contract element types with few training instances, to the extent that the best linear sliding-window classifiers are able to catch up. Nevertheless, the extra layer of BILSTM-LSTM-LR and the CRF layer of BILSTM-CRF are particularly beneficial in contract element types with few training instances, leading to significant performance improvements compared to BILSTM-LR. We can only speculate that the additional LSTM layer of BILSTM-LSTM-LR may lead to better generalization, and that the CRF layer may in effect introduce an additional training signal by allowing BILSTM-CRF to consider more directly the predicted labels of neighboring tokens.

3.2. Evaluation per Contract Element

In the second group of experiments, the methods were evaluated for their ability to identify *entire contract elements*. By contrast, the linear sliding-window classifiers and the LSTM-based methods classify individual tokens as positive or negative with respect to a particular contract element type. For the experiments of this section, each (maximal) sequence of consecutive tokens predicted to be positive with respect to a contract element type (e.g., consecutive tokens predicted to be parts of contracting parties) is treated as a single predicted contract element of the corresponding type (e.g., a single contracting party), and similarly for the gold annotations, as in our previous work [2].

For each contract element type (e.g., contracting parties), the strictest evaluation would now count as true positives only the predicted contract elements that match *exactly* gold ones, and similarly for false positives and false negatives. For example, if a method predicted "Sugar 13" to be a contracting party, missing the "Inc." of the gold "Sugar 13 Inc.", the predicted 'Sugar 13' would be a false positive and the gold "Sugar 13 Inc." would be a false negative. In practical applications, however, it often suffices to produce contract elements that are *almost* the same as the gold ones. As in our previous work, we set a threshold $t \in [0.8, 1.0]$ for each contract element type (based on requirements of our clients, the same as in the experiments of our previous work) and we consider a predicted contract element as true positive (*TP*) if (1) it is a substring of a gold contract element (of the same type) and the length of the predicted element (in characters, excluding white spaces) is at least $t\%$ of the length of the gold one, or vice versa (2) a gold contract element is a substring of the predicted one and the length of the gold element is at least $t\%$ of the length of the predicted one. Otherwise the predicted contract element is a false positive (*FP*); and the corresponding gold element is a false negative (*FN*), unless it matches another predicted element of the same type.[5] Precision, recall, and F_1 are then defined as in Section 3.1, but now using the definitions of *TP*, *FP*, *FN* of this paragraph.

Table 2 lists the results of the second group of experiments. We now include the manually crafted post-processing rules of our previous work [2] in the best linear sliding window classifiers, since they improve significantly their performance, as reported in our previous work.[6] The simplest LSTM-based method, BILSTM-LSTM-LR, equals the

[5] The values of t are: 1.0 for start, effective, termination dates; 0.9 for governing law and clause headings; 0.8 for other contract element types.

[6] The post-processing rules cannot be used when evaluating per token (Section 3.1), because they require multi-token contract elements to have been grouped into single contract elements, as in this section.

Element Type	SW-LR-ALL-POST			SW-SVM-ALL-POST			BILSTM-LR			BILSTM-LSTM-LR			BILSTM-CRF		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Title	0.84	0.93	0.88	0.83	0.93	0.88	0.96	0.94	**0.95**	0.95	**0.95**	**0.95**	**0.97**	**0.95**	**0.95**
Parties	0.96	0.85	0.90	0.95	0.85	0.89	0.98	0.89	0.93	0.96	**0.91**	**0.94**	**0.99**	0.87	0.93
Start	0.89	0.94	0.91	0.84	0.93	0.88	0.88	**0.96**	0.92	0.92	0.95	0.93	**0.93**	**0.96**	**0.94**
Effective	0.86	0.80	0.83	0.87	0.95	0.91	**0.99**	0.89	**0.94**	0.94	0.94	**0.94**	0.95	0.86	0.90
Termination	**0.79**	0.91	**0.85**	0.75	0.96	0.84	0.73	0.89	0.80	0.72	0.91	0.81	**0.74**	**0.98**	0.84
Period	**0.62**	**0.85**	**0.72**	0.51	0.80	0.63	0.40	0.71	0.51	0.44	0.75	0.55	**0.62**	0.79	0.70
Value	0.74	0.92	**0.82**	0.72	**0.94**	0.81	0.66	0.67	0.67	**0.75**	0.72	0.73	0.70	0.78	0.74
Gov. Law	**0.99**	0.93	0.96	**0.99**	0.95	0.97	0.98	**0.97**	0.97	**0.99**	**0.97**	**0.98**	**0.99**	0.96	0.97
Jurisdiction	0.99	0.75	0.85	0.98	0.78	0.87	0.90	0.85	0.88	0.92	0.82	0.86	0.93	**0.85**	**0.89**
Legisl. Refs.	**0.97**	0.88	0.92	**0.97**	0.90	**0.94**	0.86	**0.94**	0.90	0.87	0.93	0.90	0.92	**0.94**	0.93
Headings	0.94	0.80	0.86	0.94	0.80	0.86	**0.99**	**0.89**	**0.94**	**0.99**	**0.89**	**0.94**	**0.99**	0.88	**0.94**
Macro-average	0.87	0.87	0.86	0.85	**0.89**	0.86	0.85	0.87	0.86	0.86	**0.89**	0.87	**0.88**	**0.89**	**0.88**

Table 2. Precision (P), Recall (R), and F_1 score, *measured per contract element instance*. Best results per contract element type shown in bold font in gray cells.

macro-averaged F_1 score (0.86) of the best linear sliding-window classifiers (SW-LR-ALL-POST, SW-SVM-ALL-POST), *without using any manually written rules*, unlike the sliding-window classifiers that *rely extensively* on the *post-processing rules* in these experiments; without the post-processing rules, the macro-averaged F_1 score of the best linear sliding window classifiers (SW-LR-ALL, SW-SVM-ALL) drops to 0.69 (not shown in Table 2, see our previous work). This is particularly important, because the post-processing rules are very difficult to maintain in practice, since they have to be tailored to the errors of each particular classifier per contract element type.

The extra LSTM layer of BILSTM-LSTM-LR improves the macro-averaged F_1 of BILSTM-LR by one point (0.87 vs. 0.86), but BILSTM-CRF now performs even better overall (0.88 macro-averaged F_1). By looking at the F_1 scores per contract element type, we see that BILSTM-CRF now performs better or at least as well as BILSTM-LR in all but one contract element types, the exception being effective dates where BILSTM-LR is better (0.94 vs. 0.90 F_1). In several contract element types, the improvements of BILSTM-CRF compared to BILSTM-LR are very significant, with the largest improvements observed in contract periods (from 0.51 to 0.70 F_1), contract values (from 0.67 to 0.74), and termination dates (from 0.80 to 0.84). Recall that these are the three contract element types with the fewest training instances in the dataset. As in the experiments of the previous section, they are also the contract element types where all the LSTM-based methods again obtain their lowest F_1 scores, and where the linear sliding-window classifiers catch up with or exceed (in the case of SW-LR-ALL-POST) the LSTM-based methods. The extra LSTM layer of BILSTM-LSTM-LR also improves the performance of BILSTM-LR in these three contract element types (from 0.51 to 0.55, from 0.80 to 0.81, and from 0.67 to 0.72, respectively), but the improvements are smaller compared to those of BILSTM-CRF. Like BILSTM-CRF, BILSTM-LSTM-LR improves or matches the F_1 of BILSTM-LR in all but one of the eleven contract element types, the exception being jurisdiction, where BILSTM-LR is better (0.88 vs. 0.86 F_1), but the improvements are smaller compared to those of BILSTM-CRF. Overall, BILSTM-CRF appears to be better than BILSTM-LSTM-LR in the experiments of this section, in contrast to the experiments of the previous section, suggesting that although BILSTM-LSTM-LR makes fewer errors per token, the errors of BILSTM-CRF are less severe, in the sense that the thresholds t allow more of the extracted contract elements of BILSTM-CRF to be considered as successfully extracted.

4. Related Work

Our previous work [2] provides an extensive overview of related work, concluding that previous text analytics work on contracts [27,28,29] focused on classifying entire lines, sentences, or clauses, rather than extracting specific contract elements, and used much smaller datasets or fewer classes. In the broader legal text analytics context, our previous work concludes that the closest related work considered segmenting legal (mostly legislative) documents [30,31,32,33,34,35] and recognizing named entities [36,35], but the proposed methods are not directly applicable to contract element extraction; for example, they employ hand-crafted features, patterns, or lists of known entities that would have to be tailored for contracts.

More recently, Garcia-Constantino et al. [37] experimented with 97 "legal documents related to commercial law", apparently contracts or documents similar to contracts, aiming to identify the sections, subsections, appendices etc. of each document, and to extract the date of each document, the names of the parties involved, the governing law, and jurisdiction. No machine learning was involved. Instead, manually crafted pattern-matching rules were employed.

Deep learning methods have recently been successfully applied to sequence labeling tasks. For example, Ling et al. [38] used a BILSTM layer operating on characters to construct morphology-aware word embeddings, which were combined with WORD2VEC embeddings and passed on to another LSTM or BILSTM layer (with a softmax), to perform language modeling or POS tagging, respectively. Lample et al. [16] experimented with a similar method in named entity recognition, adding a CRF layer, and reporting that an alternative method that involved stacked LSTMs performed worse. Ma and Hovy [17] used a convolutional neural network (CNN) to obtain word embeddings from characters; the word embeddings were subsequently fed to a BILSTM layer followed by a CRF layer to perform POS tagging or named entity recognition. Huang et al. [15] experimented with LSTM and BILSTM layers combined with CRF layers in POS tagging, chunking, and named entity recognition. However, we are among the first to apply deep learning to legal text analytics tasks; see also [39,40].

5. Conclusions and Future Work

Building upon our previous work, we explored how deep learning methods can be used in contract element extraction. We showed that a BILSTM with a logistic regression layer (BILSTM-LR), operating on pre-trained word, POS tag, and token-shape embeddings outperforms in most cases the best methods of our previous work, which employed linear classifiers operating on fixed-size windows of tokens, without employing any manually written rules. Further improvements were observed by stacking an additional LSTM on top of the BILSTM (BILSTM-LSTM-LR) or by adding a CRF layer on top of the BILSTM (BILSTM-CRF). Experimental results indicated that BILSTM-LSTM-LR misclassifies fewer tokens, but that BILSTM-CRF performs better when methods are evaluated for their ability to extract entire contract elements. Interestingly, the additional LSTM and CRF layers were most beneficial in contract element types with few training instances.

Future work could explore if BILSTM-CRF can be improved further by using additional stacked LSTM layers, or additional BILSTM or CNN layers to produce

morphologically-aware word embeddings [38,17]. We also plan to explore data augmentation techniques [41], especially in contract element types with few training instances.

References

[1] Z. Milosevic, S. Gibson, P. F. Linington, J. Cole, and S. Kulkarni. On design and implementation of a contract monitoring facility. In *Proc. of the 1st IEEE Int. Workshop on Electronic Contracting*, pages 62–70, San Diego, CA, 2004. IEEE Computer Society Press.

[2] I. Chalkidis, I. Androutsopoulos, and A. Michos. Extracting contract elements. In *Proc. of the 16th Int. Conf. on Artificial Intelligence and Law*, pages 19–28, London, UK, 2017.

[3] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, 2011.

[4] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[5] T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proc. of the Conf. of the North American Chapter of the ACL: Human Language Technologies*, pages 746–751, Atlanta, GA, 2013.

[6] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar, 2014.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[8] Y. Goldberg. *Neural Network Methods in Natural Language Processing*. Morgan and Claypool Publishers, 2017.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[10] F. A. Gers, J. Schmidhuber, and F. A. Cummins. Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10):2451–2471, 2000.

[11] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[12] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278, Olomouc, Czech Republic, 2013.

[13] O. Irsoy and C. Cardie. Deep recursive neural networks for compositionality in language. In *Proc. of the 27th Int. Conf. on Neural Information Processing Systems*, pages 2096–2104, Montreal, Canada, 2014.

[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th Int. Conf. on Machine Learning*, pages 282–289, Williamstown, MA, 2001.

[15] Z. Huang, W. Xu, and K. Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

[16] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proc. of the Conf. of the North American Chapter of the ACL: Human Language Technologies*, pages 260–270, San Diego, California, 2016.

[17] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional LSTM-cNNs-CRF. In *Proc. of the 54th Annual Meeting of the ACL*, pages 1064–1074, Berlin, Germany, 2016.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of the 26th Int. Conf. on Neural Information Processing Systems*, Stateline, NV, 2013.

[19] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber. LSTM: A search space odyssey. *CoRR*, abs/1503.04069, 2015.

[20] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[21] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of the Int. Conf. on Artificial Intelligence and Statistics*, pages 249–256, Sardinia, Italy, 2010.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. of the Int. Conf. on Learning Representations*, San Diego, CA, 2015.

[23] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, and G. Kurian. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[24] J. Peng, L. Bo, and J. Xu. Conditional neural fields. In *Advances in Neural Information Processing Systems*, pages 1419–1427. Vancouver, Canada, 2009.

[25] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao. Recurrent conditional random field for language understanding. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 4077–4081, Florence, Italy, 2014.

[26] D. Andor, C. Alberti, D. Weiss, A. Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. In *Proc. of the 54th Annual Meeting of the ACL (long papers)*, pages 2442–2452, Berlin, Germany, 2016.

[27] M. Curtotti and E. Mccreath. Corpus based classification of text in Australian contracts. In *Proc. of the Australasian Language Technology Association Workshop*, pages 18–26, Melbourne, Australia, 2010.

[28] K. V. Indukuri and P. R. Krishna. Mining e-contract documents to classify clauses. In *Proc. of the 3rd Annual ACM Bangalore Conf.*, pages 7:1–7:5, Bangalore, India, 2010.

[29] X. Gao, M. P. Singh, and P. Mehra. Mining business contracts for service exceptions. *IEEE Transactions on Services Computing*, 5:333–344, 2012.

[30] A. Stranieri and J. Zeleznikow. *Knowledge Discovery from Legal Databases*. Springer, 2005.

[31] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. Automatic semantics extraction in law documents. In *Proc. of the 10th Int. Conf. on Artificial Intelligence and Law*, pages 133–140, Bologna, Italy, 2005.

[32] I. Hasan, J. Parapar, and R. Blanco. Segmentation of legislative documents using a domain-specific lexicon. In *Proc. of the 19th Int. Conf. on Database and Expert Systems Application*, pages 665–669, Turin, Italy, 2008.

[33] E. L. Mencia. Segmentation of legal documents. In *Proc. of the 12th Int. Conf. on Artificial Intelligence and Law*, pages 88–97, Barcelona, Spain, 2009.

[34] E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia. *Semantic Processing of Legal Texts*. Number 6036 in Lecture Notes in AI. Springer, 2010.

[35] P. Quaresma and T. Goncalves. Using linguistic information and machine learning techniques to identify entities from juridical documents. In E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, editors, *Semantic Processing of Legal Texts*, number 6036 in Lecture Notes in AI, pages 44–59. Springer, 2010.

[36] C. Dozier, R. Kondadadi, M. Light, A. Vachher, S. Veeramachaneni, and R. Wudali. Named entity recognition and resolution in legal text. In E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, editors, *Semantic Processing of Legal Texts*, number 6036 in Lecture Notes in AI, pages 27–43. Springer, 2010.

[37] M. F. Garcia-Constantino, K. Atkinson, D. Bollegala, K. Chapman, F. Coenen, C. Roberts, and K. Robson. CLIEL: Context-Based Information Extraction from Commercial Law Documents. In *Proc. of the 16th Int. Conf. on Artificial Intelligence and Law*, pages 79–87, London, UK, 2017.

[38] Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. Finding function in form: Compositional character models for open vocabulary word representation. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal, 2015.

[39] J. O' Neill, P. Buitelaar, C. Robin, and L. O' Brien. Classifying Sentential Modality in Legal Language: A Use Case in Financial Regulations, Acts and Directives. In *Proc. of the 16th Int. Conf. on Artificial Intelligence and Law*, pages 159–168, London, UK, 2017.

[40] S. N. Truong, N. Le Minh, K. Satoh, T. Satoshi, and A. Shimazu. Single and multiple layer BI-LSTM-CRF for recognizing requisite and effectuation parts in legal texts. In *Proc. of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts*, London, UK, 2017.

[41] T. Devries and G. W. Taylor. Dataset augmentation in feature space. *CoRR*, abs/1702.05538, 2017.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-165

Automatic Detection of Significant Updates in Regulatory Documents

Kartik Asooja, Oscar Ó Foghlú, Breiffni Ó Domhnaill, George Marchin, Sean McGrath
firstname.lastname@propylon.com
Propylon Ltd., Dublin 14, Ireland

Abstract. Regulations and legislations are regularly updated, which significantly burdens up the lawyers and compliance officers with a firehose of changes. However, not all changes are significant, and only a percentage of them are of legal importance. This percentage can certainly vary in different types of regulations. This paper focuses on automatic detection or ranking of meaningful legal changes, and presents a preliminary approach based on machine learning for the same, in the domain of Internal Revenue Code (IRC) related regulatory documents. Such system would provide the users with a means to quickly identify significant legal changes.

Keywords. Change detection, Version, Regulation, Regulatory Change Management, Machine Learning

1. Introduction

Lawyers, tax professionals, and compliance officers need to efficiently research and understand constantly changing regulations in order to competently respond to the updates and understand what their client/industry needs to comply with. The volume and velocity of changes and updates of laws and regulations are growing dramatically, which makes it even more difficult for professionals [5]. Following this, the legal publishers operate in a competitive, constantly changing environment where technology is supplying the unique proposition to many content products. Legal research often requires the study of the change timeline of the regulatory documents, especially in the cases of litigations, where one might need to study point-in-time changes in the regulatory framework. However, picking up the significant material or legal changes in a version history can be really expensive and cumbersome, as there can be good number of versions just accounting for changes in text formats, spellings, etc.

In this paper, we present a machine learning based approach to automatically detect the versions with significant material changes. Environmental Data and Governance Initiative (EDGI)[1] monitors government webpages to track environment related regulatory changes, and they are also doing a highly relevant project which aims at automatically identifying and prioritizing those changes[2].

[1] https://envirodatagov.org/website-monitoring/
[2] https://github.com/edgi-govdata-archiving/web-monitoring

2. TimeArc

Propylon's[3] TimeArc[4] platform is a legal research software solution that enables easy understanding of changes in legislation and regulation with the help of version timeline, with redlining comparison capabilities and point-in-time hyperlinking. It enables you to see the most up-to-date information available as well as a historical view, with access to a full revision history of every change ever made to any given document. This allows the user to find and compare historical changes, discover intent, and filter by commentary, context, and editorial overview. All of the information is available in an easy-to-use, efficient, browser-based tool that gives new and enhanced insights, as shown in Figure 1.

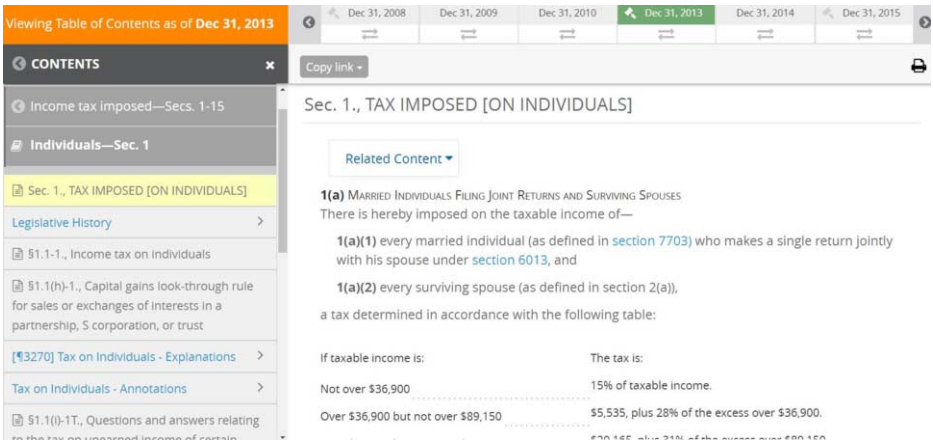


Figure 1. Sample Document Versions on TimeArc Platform. Versions with *gavel* show the significant ones.

However, TimeArc is designed to pick up even the smallest of changes in documentation and present them on the timeline in order to ensure that strict audit trails are maintained. This can result in the plotting of potentially insignificant changes on document timelines. Therefore, in this work, we focus on highlighting the significant changes in the document version timeline.

.01 Amended by P.L. 113-295 (deadwood amendment), P.L. 112-240, P.L. 110-185 (conforming amendment), P.L. 110-28, P.L. 109-222, P.L. 108-357 (conforming amendments), P.L. 108-311, P.L. 108-27, P.L. 107-16, P.L. 106-554 (conforming amendment), P.L. 105-277 (technical corrections), P.L. 105-206, P.L. 105-34, P.L. 104-188, P.L. 103-66, P.L. 101-508, P.L. 101-239, P.L. 100-647, P.L. 99-514, P.L. 97-448 (clarification, not an amendment), P.L. 97-34, P.L. 95-600, P.L. 95-30, P.L. 91-172, P.L. 89-809 and P.L. 88-272. For details, see the Code Volumes.

Figure 2. Sample 1: Human commentary section for a document

[3]https://www.propylon.com/
[4]https://www.propylon.com/legal-research/

.01 Historical Comment: Proposed 7/13/55. Adopted 2/3/56 by T.D. 6161. Amended 5/24/71 by T.D. 7117, 12/20/74 by T.D. 7332, and 4/4/2008 by T.D. 9391. [Reg. §1.1-1 does not reflect P.L. 95-600 (1978), P.L. 97-34 (1981), P.L. 97-488 (1983), P.L. 99-514 (1986), P.L. 100-647 (1988), P.L. 103-66 (1993), P.L. 107-16 (2001), P.L. 108-27 (2003), P.L.108-311 (2003) or P.L. 112-240 (2013). See ¶3260.045 et seq. and ¶3270.01].

Figure 3. Sample 2: Human commentary section for a document

3. Data

Definition of a significant change in a document can be highly contextual depending on the user and domain. Our use case deals with tax professionals and documents related to the US IRC and Treasury Regulations, provided by a legal publisher. The documents have a parallel human commentary, as shown in the figures 2 and 3. It summarizes the changes in a version for a section of data by giving citations to related Public Laws (P.L.) and Treasury Decisions (T.D.). Based on the suggestions from subject matter experts, we consider a change in a document as significant if there is a change in the citations within this commentary, especially to the ones related to Public Laws and Treasury Decisions. This implies that if there is a relevant regulatory change in P.L./T.D.s, it requires the publisher to make major amendments in related regulatory documents. However, still there can be some potential significant updates to the regulatory documents, which are not dependent on the changes in the P.L./T.D.s. For the experiments reported here, we do not consider these versions as significant, as it would require expensive tagging by subject matter experts.

We use the documents from 2005 to 2015 for our experiments. Total number of versions over all the documents in the data is 41,965, out of which only 24,839 ($\sim$ 59%) are significant, as per the above definition.

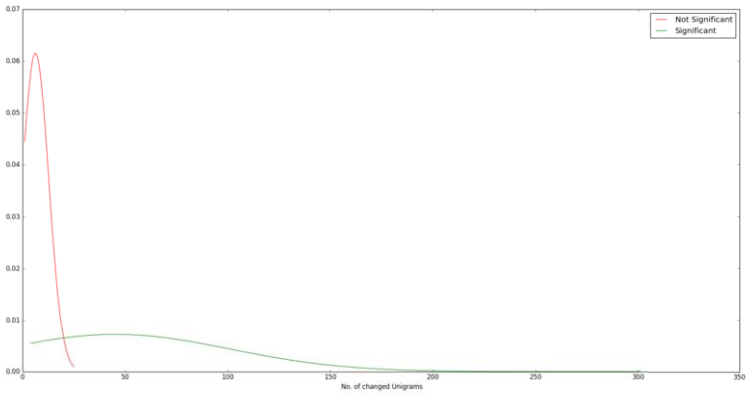


Figure 4. Data distribution (assuming Normal) against the number of unigrams in the symmetric difference, after removing 5% tail outliers.

4. Learning to Identify Significant Changes

As we have human provided commentary on the versions for only a section of our overall data from the client, the aim of this work is to learn and evaluate a machine learning based approach to automatically identify significant legal material changes for other documents in the same domain. This can be considered as a binary text classification problem, where the classes are significant change (positive class) and insignificant change (negative class).

We assume D_t^i and D_{t+1}^i represent the bag of words (BoW) sets present in the i-th document at time t and t+1 respectively. We use the symmetric difference between the BoW sets of the documents at consecutive time steps to define a single data instance, thus considering the correlation between the added and deleted words to the significance of a document revision. Therefore, the training data instance takes the following form: $(D_t^i \triangle D_{t+1}^i, y)$, where, $D_t^i \triangle D_{t+1}^i$ represents the symmetric difference between the versions, and y represents a boolean label for the class.

We consider unigrams and their counts as the features, and use Support Vector Machines (SVM) classification algorithm. Two methods were employed to evaluate our classifier: 10 fold cross validation, and 70:30 split of the total data as train and test datasets. We employ the LibSVM library [1] for SVM classifier using Weka machine learning toolkit [2]. The parameters for SVM classifier are as follows: SVM type = C-SVC, kernel function = Radial Basis Function. Features are ranked using the feature selection algorithm InfoGain [3]. Table 1 summarizes the results for the identification of the positive class. We can see that just using the count of unigrams as features can model a good predictor of the significant changes. This follows the data distribution graph shown in the figure 4, implying that if there are many added or deleted unigrams in a version, it leads to a significant version. Moreover, with unigram features, the classification improves significantly in comparison to just using the counts.

Feature	**Precision**		**Recall**		**F-measure**	
	10-fold	Train-test	10-fold	Train-test	10-fold	Train-test
Unigrams	0.925	0.912	0.910	0.890	0.917	0.901
Unigrams changed count	0.738	0.734	0.729	0.725	0.731	0.727

Table 1. Classification performance (weighted avg. metrics)

5. Conclusion

In this work, we present a preliminary approach to automatically mine the significant versions in a document timeline. Initial results from the classifier show a good performance, which can clearly enable the user to easily focus on the significant changes. The significant versions in the document timeline are highlighted in the TimeArc platform allowing the users to quickly navigate between meaningful changes in the law without seeing editorial, typographical or stylistic changes to content, as shown in demo in the figure 1. As future work, we would work on improving the classification performance by identifying more features for this problem, and by using deep learning algorithms.

References

[1] Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), p.27.

[2] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), pp.10-18.

[3] Mitchell, T.M., 1997. Machine learning. WCB.

[4] Akbani, R., Kwek, S. and Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. Machine learning: ECML 2004, pp.39-50.

[5] Asooja, K., Bordea, G., Vulcu, G., OBrien, L., Espinoza, A., Abi-Lahoud, E., Buitelaar, P. and Butler, T., 2015. Semantic annotation of finance regulatory text using multilabel classification. LeDA-SWAn (to appear, 2015).

This page intentionally left blank

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-171

A Computational Model of Moral and Legal Responsibility via Simplicity Theory

Giovanni SILENO [a,1], Antoine SAILLENFEST [b] and Jean-Louis DESSALLES [a]
[a] *LTCI, Télécom ParisTech, Université Paris-Saclay, 46 rue Barrault, Paris, France*
[b] *Geronimo Agency, 33 rue d'Artois, Paris, France*

Abstract. Responsibility, as referred to in everyday life, as explored in moral philosophy and debated in jurisprudence, is a multiform, ill-defined but inescapable notion for reasoning about actions. Its presence in all social constructs suggests the existence of an underlying cognitive base. Following this hypothesis, and building upon *simplicity theory*, the paper proposes a novel computational approach.

Keywords. moral responsibility, legal responsibility, simplicity theory, foreeseability, inadvertence, risk, negligence

1. Introduction

The notion of *individual responsibility* is paramount in informal social relationships as much as in formal legal institutions. With the (supposedly) near advent of autonomous entities, its formalization becomes a pressing problem. In human societies, responsibility attribution is a spontaneous and seemingly universal behaviour. Non related ancient legal systems (e.g. [9]) bear much resemblance to modern law and seem perfectly sensible nowadays. This universality suggests that responsibility attribution may be controlled, at least in part, by fundamental cognitive mechanisms. Experimental studies showed that various parameters influence moral responsibility attribution [15]. For instance, people are more prone to blame (praise) an agent for an action if they are closer to the victims (beneficiaries), if the outcome follows in a simple way from the action or if the agent was able to foresee the outcome. Several of these parameters, such as the agent's foreseeing ability, are purely cognitive. Theories of law indeed take cognition into account with notions such as *mens rea*. Following this idea, the present paper attempts to bridge the gap between cognitive modelling and theory of law.

The AI & Law literature proposes two main approaches to *responsibility attribution*. The structural approach attempts to capture reasoning constructs using ontologies [10], inference [14] or stories [1]. The probabilistic approach focuses on quantifying the relative support of evidence in the reasoning process, e.g. via Bayesian inference [6] or causal Bayesian networks [7,3]. Hybrid proposals exist as well [17]. The present work introduces an alternative framework, using notions from *simplicity theory* [4], offering a potential ground for unification: because *simplicity theory* relies on the computation of Kolmogorov-like complexities, it involves both structural and quantitative aspects.

The paper proceeds as follows. In § 2, we consider a few accounts of the notion of responsibility with some case examples. In § 3, we briefly introduce *simplicity theory* and show how it can deal with moral evaluation. In § 4, we evaluate it based on the given case examples. A note on further developments ends the paper.

[1]Corresponding author: giovanni.sileno@telecom-paristech.fr

2. Causal, Legal and Moral Responsibilities

For some legal scholars, a theory of responsibility should rely on legal causation rather than on factual causation (see the overviews on causal minimalism given in [10], [12]). Indeed, the Greek word for "cause" started as a legal term [11]. Consider this real case:

Example 1. (Two bad hunters) *Two hunters negligently fired their shotguns in the direction of their guide, and a pellet lodged in his eye. Because it was impossible to tell which hunter fired the shot that caused the injury, the court held both hunters liable.*[2]

Here, one of the two hunters is held responsible despite the fact that he did not materially cause the damage. Physical causation is rarely matter of dispute, and when it is, as in the previous case, it is irrelevant to formulate a legal judgement. By contrast, legal causation is always relevant and is much debated when attributing responsibility (with variations depending on the different legal traditions). Consider the following case:

Example 2. (Navigating oil) *At a landing stage, furnace oil spilled into the water for defendants' negligence. The oil spread on the water surface, reaching a nearby ship on which welding work was being carried out. Sparks ignited the oil, which caught on fire damaging several vessels. The court held that contamination damage caused by the oil was reasonably foreseeable, but that damage caused by fire was not foreseeable and was thus too remote for recovery.*[3]

The core of the dispute was to settle on *foreseeability*, i.e. the ability to predict the consequences of an event or action. Beyond foreseeability, events would be too *remote* to the defendant to be accounted liable for, even if they were enabling the actual chain of causation. Although foreseeability is a fictional device, knowing what-caused-what or what-enabled-what—*pace* causal minimalists—influences its evaluation:

Example 3. (Navigating oil, cont'd) *Further evidence revealed the presence of floating flammable objects in the water which, combined with the oil, made the lightning of the fire more probable. The court held the defendant liable, because, seen the magnitude of the risk, a reasonable person would have reacted to prevent it.*[4]

The second judgement not only considers the ability to foresee alternative causal chains, but also takes the magnitude of the risk into account. However, not every responsibility attribution is about the agents' rational abilities. Consider this simple case:

Example 4. (A broken vase) *A person enters in a shop and breaks inadvertently a vase. According to the law, she is usually liable to provide compensation, but not to be blamed.*

Even when people are making reasonable choices, things may go wrong. These cases are usually under the scope of law (but not necessarily of morality), in order to apply a fairer redistribution of the losses amongst the parties (principle of *equity*).

Legal Responsibility and its Boundaries Legal systems usually have distinct mechanisms to decide on liability (*who has to provide remedy?*, as in the previous examples) and on blame (*who has to be punished?*). In general, guiltiness is attributed by proving a combination of factual elements under the scope of law (*actus reus*) and mental elements relevant to the case (*mens rea*). Consider however this famous paradox [13, Ch. 10]:

[2] Summers v. Tice (1948), 33 Cal.2d 80, 199 P.2d 1.
[3] Overseas Tankship (UK) Ltd v. Morts Dock and Eng. Co Ltd or "*Wagon Mound (No. 1)*" (1961), UKPC 2.
[4] Overseas Tankship (UK) Ltd v The Miller Steamship Co or "*Wagon Mound (No. 2)*" (1967), 1 AC 617.

Example 5. (The desert traveller). *A desert traveller T has two enemies. Enemy 1 poisons T's canteen and Enemy 2, unaware of Enemy 1's action, empties the canteen. A week later, T is found dead and the two enemies confess to action and intention. It is then discovered that T never drank from the canteen and died by dehydration.*

From a causal point of view, this example contains a *pre-emption*: an event prevents another event from being successful. Is Enemy 1 guilty? In principle, law disregards potential outcomes, so the answer is no. Intuitively, however, Enemy 1 is morally guilty. And many legal systems do attribute some charge to the offender who willingly initiated a course of action that may have lead to a crime (e.g. *attempted murder*).

3. Theoretical Framework

This section briefly presents *simplicity theory* (ST) as a theoretical basis to construct computational models of judgement. ST is a cognitive theory stemming from the observation that human individuals are highly sensitive to *complexity drops* [4]: i.e. to situations that are *simpler to describe than to explain*. The theory builds on notions and tools from *algorithmic information theory* (AIT) that are redefined with respect to cognitive agents. It has been used to make predictions, confirmed empirically, about what humans would regard as *unexpected*, *improbable*, and *interesting* [5,15,16].[5]

Unexpectedness A central notion in ST is *unexpectedness* (U), defined as:

$$U(s) = C_W(s) - C_D(s) \tag{1}$$

where s is a situation, $C_W(s)$ is the complexity of the circumstances that were necessary to generate s, $C_D(s)$ is the complexity of describing s. The two complexities are versions of *Kolmogorov's complexity*, which, informally, is the length in bits of the shortest description of an object. ST distinguishes *causal* complexity ($C_W(s)$) from the usual *description* complexity ($C_D(s)$). Determining a causal path requires adding the complexities of making a choice at successive choice points. If there are k equivalent options at a choice point, one needs $\log_2(k)$ bits to make a decision. On many occasions, $C_W(s)$ corresponds to the logarithm of the probability of occurrence. Complexity computations, however, have a broader range of applicability, as for instance when dealing with unique events. Using $C_W(s)$ we can define the *causal contribution* of a situation s_1 to bringing about a second situation s_2:

$$R(s_1, s_2) = C_W(s_2) - C_W(s_2||s_1) \tag{2}$$

where $C_W(s_2||s_1)$ is the complexity of causally generating s_2, starting from a state of the world in which s_1 holds. If $R(s_1, s_2) = 0$, the two events are independent. If $R(s_1, s_2) > 0$ (respectively < 0), s_1 concurs positively (negatively) to the occurrence of s_2.

The description complexity $C_D(s)$ specifies the shortest *determination* of an object s. For instance, the shortest determination of s may consist in merely retrieving it from memory (think of referring to famous people). In this case, $C_D(s)$ amounts to the complexity of the parameter controlling the retrieval, i.e., considering memory as an ordered set, the $\log_2$ of the index of the object in that set (frequently used objects have smaller indexes). Applying similar considerations to spatio-temporal properties, we observe that C_D captures the distance (as inverse of proximity) of the agent to the situation.

[5] For a general presentation see: `http://simplicitytheory.org`.

Points of View For any agent A, C_W^A will denote the generation complexity computed by A using her knowledge. Different *points of view* may lead to alternative computations of causal complexity for the same situation.

Emotion and Intention Unexpectedness captures the epistemic side of a *relevant* experience. For the *epithymic* (i.e. concerning desires) side, ST refers to a representation of *emotion* limited to considering intensity E and valence ε. Focusing only on intensity, we define the *actualized* (or *hypothetical*) *emotion* as $E_h(s) = E(s) - U(s)$, pruning the emotion of its unexpectedness[6]. Intention is driven by E_h^A, computed from the point of view of an agent A who considers performing action a. If A sees a as the shortest causal path to s, $U^A(s) = U^A(a) + U^A(s||a)$, and intention turns out to be:

$$I(a) = E^A(s) - U^A(s||a) - U^A(a) \tag{3}$$

When a is intended (volitional), $U(a) = 0$. This term, when non-zero, represents ***inadvertence***. Note that in the more general case, intention should result from an aggregation of similar components for different outcomes s_i.

Moral Responsibility and Judgement Our central claim is that the difference between intention and of moral responsibility is one of *point of views*. To obtain intention, we consider the point of view of the actor A for all the components. When performing moral evaluation, however, the observer applies her own point of view (we omit superscript O), except for the elements concerning the action, which are computed using her *model* of the actor. The ***moral responsibility*** M attributed to A by observer O is defined as:

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a) \tag{4}$$

The superscript $^{\downarrow A}$ means that O uses her model of A to compute U (e.g. a prescribed role, a reasonable standard, etc.). If we introduce the actualized emotion term we have: $M(a) = E_h(s) + U(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$, from which, making C_W and C_D explicit, we can extract the ***causal responsibility*** component:

$$R^{\downarrow A}(a,s) = C_W(s) - C_W^{\downarrow A}(s||a) \tag{5}$$

This formula captures how much A's action a was supposed to bring about s in A's mind. If we suppose than $C_D^{\downarrow A}(s|a) \approx 0$ — a simplification possible when the conceptual relation between cause and effect is proximate (i.e. in A's model, the action is directly linked to the outcome) — the resulting equation is:

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a) \tag{6}$$

In words, the intensity of moral evaluation increases with the *actualized emotional intensity* and with *causal responsibility*, decreases with the *remoteness* of the consequence to the observer (proximate situations are simpler to describe) and with *inadvertence*.[7]

Now, imagine the case of a famous singer who is killed as a casual bystander in a car accident. The popular emotion might be so strong that the police have to save the car

[6] In a utilitarian perspective, E_h may be interpreted as the logarithmic version of the expected value, and E as the logarithm of the absolute value of gain or loss.

[7] Like for intention, a complete *moral judgement* of a positive action a should take into account also the evaluation of its *omission*, in order to capture e.g. the fact that someone may act negatively to avoid even worst consequences (cf. attenuating circumstances).

driver from being lynched. An impartial judge must consider the victim as if she were any person. This means that *equality* in judgement is obtained by reducing the impact of C_D, i.e. by *recomplexifying* the mental simplification due to proximity effects.

4. Applying Simplicity Theory to Judgement

We now examine how the framework presented above matches our examples.

Two bad hunters Two hunters (A_1, A_2) fire *negligently* at their guide (a_1, a_2), resulting in his injury (s). Causal contributions—$R(a_1,s)$ or $R(a_2,s)$—cannot be determined. Negligence is captured when actors fail to *foresee* the unlawful consequences of their action: $C_W^{A_1}(s||a_1) = C_W^{A_2}(s||a_2) \gg 0$. However, it is reasonable to expect that the two actions may have resulted in that outcome (note that $C_W(s) \gg 0$): $C_W^{\downarrow A_1}(s||a_1) = C_W^{\downarrow A_2}(s||a_2) > 0$ and $R^{\downarrow A_1}(a_1,s) = R^{\downarrow A_2}(a_2,s) > 0$. Therefore, both hunters receive the same moral evaluation. Generalizing this case, the ***negligence*** of an actor A for an action a w.r.t. a consequence s is defined as:

$$N^A(a,s) = C_W^A(s||a) - C_W^{\downarrow A}(s||a) \tag{7}$$

Navigating oil The oil leakage at the landing stage (s_1) results from an omission of adequate care ($a = \neg b$) by defendant A. The case centers around responsibility attribution for the fire at the near wharf (s_2). The court held that though s_1 was foreseeable, s_2 was not: $C_W^{\downarrow A}(s_1||a) \sim 0$ and $C_W^{\downarrow A}(s_2||s_1) \gg 0$, and $R(s_1,s_2) \sim 0$. Integrating the C_D terms, we define A's ***foreseeability*** of the consequence s of an action a as negated unexpectedness:

$$F^A(a,s) = -U^{\downarrow A}(s||a) \tag{8}$$

(F^A is in $]-\infty,0]$, 2^{F^A} in $[0,1]$; $F^A = 0$, $2^{F^A} = 1$ when s is perfectly foreseeable after a.)

Navigating oil, cont'd Due to the presence of flammable objects (s_1'), the defendant should have reasonably anticipated the consequences: $C_W^{\downarrow A}(s_2||a \wedge s_1) > C_W^{\downarrow A}(s_2||a \wedge s_1 \wedge s_1')$. Foreeseability increases, and so does responsibility. The court made also an argument about weighting of risks. Traditionally, risks are approached with *expected value*. Considering $E(s)$ as the "win" value (loss in this case), the ***risk*** can be defined as:

$$K^A(a,s) = E(s) - U^{\downarrow A}(s||a) = E(s) + F^A(a,s) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) \tag{9}$$

This view agrees with Hart and Honoré's [8] consideration of risk as a generalization of foreseeability, providing an *upper bound* for the damages to be paid.

A broken vase A person A slips in a shop (a) and breaks a vase (s). For a person to slip is unexpected but still possible: $U(a) > 0$ with a good probability of breaking something ($U^A(s||a) \sim 0$, $C_W(s||a) > 0$ and $R(a,s) \gg 0$). We get: $M(a) \approx E(s) - U^{\downarrow A}(a)$. This expression accounts for the fact that the agent and the shopkeeper may have different evaluations of $M(a)$, due to their different appraisal of $E(s)$.

The desert traveller Enemy 1 (E_1) poisons the canteen (a_1); Enemy 2 (E_2) empties the canteen (a_2). Instead of getting poisoned (s_1), the desert traveller gets dehydrated (s_2) and dies (s). We have: $C_W(s) \gg 0$, $C_W(s_1||a_1) = C_W(s_2||a_2) = C_W(s||s_2) = C_W(s||s_1) = 0$, and $C_W(s_2||a_1) \gg C_W(s_2||a_2) = 0$. Then, $R(a_1,s_2) = 0$, but also $C_W(s||s_2) - C_W(s||s_2 \wedge a_1) = 0$, which explains why E_1 is not judged causally responsible for the occurrence of s, knowing that s_2 was the case. However, $R^{\downarrow E1}(a_1,s) \gg 0$, which explains why E_1 is regarded as morally responsible (Eq. (6)).

5. Conclusion and Further Developments

The hypothesis advanced here is that moral and legal responsibility attributions share a fundamentally similar cognitive architecture. We could derive from *simplicity theory* formal definitions of: *intention* (3), *moral responsibility* (4, 6), *causal responsibility* (5), *inadvertence*, *negligence* (7), *foreseeability* (8), *risk* (9). These results are however preliminary, and further investigation is needed to compare them with existing proposals (see § 1). For instance, the analytic definitions of *degree of responsibility* and *blame* given in [7] are aligned with those of *causal contribution* (2) and *causal responsibility* (5).

As observed in the domain of legal ontologies [2], legal reasoning builds upon *normative knowledge* (qualifying behaviour as allowed and disallowed) and *responsibility knowledge* (assigning responsibility for the behaviour). The former is fed mostly by world definitional knowledge, the second by world causal knowledge. Our model is aligned with this analysis, for the crucial role of world complexity (C_W). For its cognitive flavour, our proposal offers an alternative contribution on responsibility in the field of AI and Law. Furthermore, for its grounding on Kolmogorov complexity, it offers a computational alternative to probability-based approaches (e.g. [6]), not requiring the reference to *a priori* probabilities, but referring to cognitively grounded elements. The richness of the framework opens new spaces for further interaction with legal analysis, analytic proposals, and for comparisons with empirical results.

References

[1] F. J. Bex, P. J. Van Koppen, H. Prakken, and B. Verheij. A hybrid formal theory of arguments, stories and criminal evidence. *Artificial Intelligence and Law*, 18(2):123–152, 2010.

[2] J. A. Breuker and R. G. F. Winkels. Use and Reuse of Ontologies in Legal Knowledge Engineering and Information Management. *Proc. of Int. Workshop on Legal Ontologies (LegOnt'03)*, 2003.

[3] H. Chockler, N. Fenton, J. Keppens, and D. A. Lagnado. Causal analysis for attributing responsibility in legal cases. *Proc. of 15th Int. Conf. on Artificial Intelligence and Law (ICAIL15)*, pages 33–42, 2015.

[4] J. L. Dessalles. Algorithmic simplicity and relevance. *Algorithmic probability and friends*, 7070 LNAI:119–130, 2013.

[5] A. Dimulescu and J.-L. Dessalles. Understanding Narrative Interest : Some Evidence on the Role of Unexpectedness. *Proc. of 31st Conf. of the Cognitive Science Society*, pages 1734–1739, 2009.

[6] N. Fenton, M. Neil, and D. a. Lagnado. A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, 37(1):61–102, 2012.

[7] J. Y. Halpern. Cause, responsibility and blame: A structural-model approach. *Law, Probability and Risk*, 14(2):91–118, 2015.

[8] H. L. A. Hart and T. Honoré. *Causation in the Law*. Clarendon Press, 1985.

[9] U. Lau and T. Staack. *Legal Practice in the Formative Stages of the Chinese Empire*. Brill, 2016.

[10] J. Lehmann, J. A. Breuker, and P. W. Brouwer. Causation in AI & Law. *Artificial Intelligence and Law*, 12(4):279–315, 2004.

[11] R. McKeon. The Development and The Significance of the Concept of Responsibility. *Revue Int.e De Philosophie*, 11(39):3–32, 1957.

[12] M. S. Moore. *Causation and Responsibility*. Oxford University Press, 2009.

[13] J. Pearl. *Causality*. Cambridge University Press, 2009.

[14] H. Prakken. An exercise in formalising teleological case-based reasoning. *Artificial Intelligence and Law*, pages 49–57, 2002.

[15] A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. *Proc. of 34th Conf. of the Cognitive Science Society*, pages 947–952, 2012.

[16] A. Saillenfest and J.-L. Dessalles. Some Probability Judgments may Rely on Complexity Assessments. *Proc. of 37th Conf. of the Cognitive Science Society*, pages 2069–2074, 2015.

[17] B. Verheij. To catch a thief with and without numbers: Arguments, scenarios and probabilities in evidential reasoning. *Law, Probability and Risk*, 13(3-4):307–325, 2014.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-177

Toward Linking Heterogenous References in Czech Court Decisions to Content

Jakub HARAŠTA [a,1], Jaromír ŠAVELKA [b]
[a] *Institute of Law and Technology, Faculty of Law, Masaryk University, Czechia*
[b] *Graduate Student, Intelligent Systems Program, University of Pittsburgh, USA*

Abstract. In this paper we present initial results from our effort to automatically detect references in decisions of the courts in the Czech Republic and link these references to their content. We focus on references to case-law and legal literature. To deal with wide variety in how references are expressed we use a novel distributed approach to reference recognition. Instead of attempting to recognize the references as a whole we focus on their lower level constituents. We assembled a corpus of 350 decisions and annotated it with more than 50,000 annotations corresponding to different reference constituents. Here we present our first attempt to detect these constituents automatically.

Keywords. case law analysis, reference recognition, conditional random fields, information extraction

1. Introduction and Challenge

Information extraction from unstructured (textual) data such as court decisions is challenging. References to other documents provide a rich set of information that could be useful in many practical applications, especially in legal information retrieval (IR). In our work we assess the possibility of recognizing the references automatically. We focus on references to case-law and scholarly literature. We intentionally leave aside references to statutory law and regulations because these appear to be quite uniform and significantly less challenging. In addition to detecting references we explore the possibility of detecting the piece of content they relate to in a decision (often quotation or a paraphrase from the referred document).

In the Czech Republic the courts do not observe a single citation standard such as the Bluebook in the United States. Instead there are multiple court-specific standards. On top of that the standards are not strictly enforced and they are subject to changes. Consider the following example of three different references that all refer to the same decision:[2]

> Decision of Constitutional Court of the Czech Republic Pl. ÚS 4/94 published in Collection of Decisions and Decrees of the Constitutional Court of the Czech Republic, year 1994, no. 46.

> Docket no. Pl. ÚS 4/94, Collection of decisions, volume 2, decision no. 46, published as no. 214/1994 Sb.

[1] Corresponding Author: jakub.harasta@law.muni.cz.

[2] All of the references are referring to an important decision of the Constitutional court that established test of proportionality within the Czech law. All references are translated to English.

Decision of October 12, 1994, docket no. Pl. ÚS 4/94, N 46/2 SbNU 57, 214/1994 Sb.

As can be seen the identifiers include docket numbers, vendor-specific identifiers, court reports or ID under which a case is listed in a specific legal information database.

2. Related Work

Significant amount of work was done in the area of reference recognition for purpose of bringing references under a set of common standards (for use in Italian legislation see [6]) or to account for multiple variants of the same reference and vendor-specific identifiers (see [5]). Both [6] and [5] are based on use of regular expressions. Language specific (Czech) work in [2] focused on detecting and classifying references to other court decisions and acts.

Our work is also focused on the content carried by a reference. Content of a reference is usually ignored; with exception of [8], [9], and [7]. [8] allows to determine which sentences near a reference are the best ones to represent the Reason for Citing. [9] uses metadata obtained by [8] that allow to explore so called Reason for Citing to create semantic-based network. [7] uses manual annotation with subsequent automated reference recognition and detection of topics of paragraphs using GATE framework [1].

3. Task

3.1. Specification

Because of the lack of a single citation standard we decided to understand references as consisting of smaller units. The smaller units are more uniform and therefore better suited for automatic detection. Some references may contain many of these units whereas other references may only have some of them. The units may appear in almost any order within a single reference.

For references to case-law the following constituents were identified:

- **c:id** - a unique court decision identifier,
- **c:court** - the court that issued the referred decision,
- **c:date** - the date on which the decision was issued,
- **c:type** - the type of the decision (e.g., decision, decree, opinion).

References to scholarly literature consist of these elements:

- **l:title** - the title of the referred work,
- **l:author** - the author or multiple authors of the referred work,
- **l:other** - other information of interest, such as place or year of publication.

Both types of references may also contain the following elements:

- **POI** - a pointer to a specific place in the decision or literary work (e.g., a page),
- **content** - the content associated with the reference (e.g., quotation, paraphrase).

References can also be expressed implicitly. In this way the courts usually refer decisions or scholarly literature that have been referred earlier in the decision. Since this occurs quite often we have created a special **implicit** constituent.

It is possible to encounter two tests for assessing impartiality of judge even in the case-law of *European Court of Human Rights*: subjective test stemming from personal convictions of judge deciding given case, objective test following whether sufficient assurances exist to exclude any legitimate doubt (comp. decision in **case of Saraiva de Carvalho v. Portugal,** decided on April 22, 1994**, application no. 15651/89**, eventually **Gautrin and others v. France**, decided on May 20, 1998, **application no. 21257/93**). In this context, it is worth mentioning that the Constitutional court noted that procedure of exclusion of judge from deciding upon case is one of the procedural guarantees of impartiality of court. When assessing whether the objective aspect of doubt on impartiality is present, even appearance may play a role in this regard [eg. decision of the *European Court of Human Rights* **Piersack v. Belgium,** decided on October 1, 1982**, application no. 8692/79,** §30; compare also decision **Wettstein v. Switzerland** decided on December 21, 2000**, application no. 33958/96,** §42 – 44, and decision of the *Constitutional Court* **docket no. III. ÚS 441/04** decided on January 12, 2005, published as **N 6/36 SbNU 53**].

Figure 1. Sample of annotated decision. Types of annotations are as follow: *c:court*, content, **c:id**, c:type, c:date, POI

	c:id	c:type	c:date	c:court	l:author	l:title	l:other	POI	implicit	content
annotations count	12043	5964	5449	4305	3426	2406	2609	3760	1202	10129
average per doc	34.41	17.04	15.57	12.30	9.79	6.87	7.45	10.74	3.43	28.94
gold count	6237	2992	2687	2236	1863	1176	1251	1854	483	4903
average gold per doc	17.82	8.55	7.68	6.39	5.32	3.36	3.57	5.30	1.38	14.01
strict agreement (inter)	70.62	80.72	86.27	73.61	73.88	50.25	44.92	73.56	27.62	26.51
overlap agreement (inter)	81.08	84.10	88.31	77.03	83.22	69.12	70.37	80.45	38.60	59.54
agreement (gold)	80.36	87.58	91.22	82.74	81.88	57.61	59.06	79.25	41.64	32.96

Table 1. Summary statistics of the data set.

3.2. Data Set

The data set consists of 350 decisions of the top-tier courts in the Czech Republic (160 Supreme Court, 115 Supreme Administrative Court, 75 Constitutional Court).[3] The shortest decision has 4,746 characters whereas the longest decision has 537,470 characters (average 36,148.68).

Decisions were annotated by thirteen annotators who were paid for their work. The annotators were trained to follow the annotation manual by means of dummy runs (i.e., annotation of documents that are not included in the data set). To ensure high quality of the resulting gold data set the three most knowledgeable annotators were appointed curators of the data set. Each document was then further processed by one of the curators. A curator could not be assigned a document that he himself annotated. The goal of the curators was to evaluate correctness of each annotation and to fill-in missing annotations. The result of their work is the gold data set.

The annotators generated 51,293 annotations (i.e., approximately 146.6 annotations per document). The detailed counts are shown in the first two rows of Table 1. The numbers correspond to all the annotations where each document was processed by two annotators. The second (gold count) and the third (avg gold per doc) rows of Table 1 provide details of the gold data set created by the curators. These entries do not contain duplicate annotations as opposed to the first two rows.

[3]The decisions were downloaded from publicly available online databases with exception of 8 cases. These were unavailable from public database of respective court and were retrieved from commercial information systems.

We report three types of inter-annotator agreement in the bottom three rows of Table 1. The strict agreement is the percentage of the annotations where the annotators agree exactly (i.e., the start and end character offsets are the same). The overlap agreement relaxes the exact matching condition—it is sufficient if the two annotations overlap by at least one character. The agreement (gold) reports the percentage of the annotations that were evaluated as correct by the curators.

3.3. Detecting Reference Constituents Automatically

We attempted to recognize the constituents of references automatically. This corresponds to detecting text spans representing the types described in Section 3.1 and summarized in Table 1. As a prediction model we use conditional random fields (CRF).[4] A CRF is a random field model that is globally conditioned on an observation sequence O. The states of the model correspond to event labels E. We use a first-order CRF in our experiments (observation O_i is associated with E_i). [3,4] We train a CRF model for each of the 10 labels. Although this is certainly suboptimal, we use the same training strategy and features for all the models. We reserve fine-tuning of models for future work.

In tokenization we consider an individual token to be any consecutive sequence of either letters, numbers or whitespace. Each character that does not belong to any of these constitutes a single token. Each of the tokens is then a data point in a sequence a CRF model operates on. Each token is represented by a small set of relatively simple features. Specifically, the set includes:

- *position* – position of a token within a document.
- *lower* – a token in lower case.
- *stem* and *aggressive stem* – two types of token stems.[5]
- *sig* – a feature representing a signature of a token.
- *length* – token's length.
- *islower* – true if all the token characters are in lower case.
- *isupper* – true if all the token characters are in upper case.
- *istitle* – true if only the first of the token characters is in upper case.
- *isdigit* – true if all the token characters are digits.
- *isspace* – true if all the token characters are whitespace.

For each token we also include *lower*, *stem* and *aggressive stem*, *sig*, *islower*, *isupper*, *istitle*, *isdigit*, and *isspace* features from the five preceding and five following tokens. If one of these tokens falls beyond the document boundaries we signal this by including *BOS* (beginning of sequence) and *EOS* (end of sequence) features.

4. Results and Discussion

4.1. Results

To evaluate the performance we use a 10-fold cross-validation. Table 2 summarizes the results of the experiments. The first two rows report the number (and average per docu-

[4] We use the CRFSuite which is available at `www.chokkan.org/software/crfsuite/`

[5] A stemmer for Czech implemented in Python by Luís Gomes was used for stemming. The stemmer is available at `http://research.variancia.com/czech_stemmer/`

	c:id	c:type	c:date	c:court	l:author	l:title	l:other	POI	implicit
predicted count	5891	2967	3001	1936	1454	786	909	1779	158
average per doc	16.83	8.48	8.57	5.53	4.15	2.25	2.60	5.08	0.45
strict agreement (gold)	65.22	75.95	75.00	56.81	64.88	43.22	49.07	70.02	6.86
overlap agreement (gold)	70.86	78.40	75.14	57.86	74.77	58.31	61.20	75.03	21.53

Table 2. Results of automatic detection of reference constituents

ment) of annotations of each type that were automatically generated for the whole data set. The third row reports the agreement with the gold standard where the equality of annotations is measured strictly (i.e., the start and end offsets both need to match exactly). The fourth row reports the agreement where the annotations to be considered equal just need to overlap by at least one character.

As one would expect the performance of the models correlates to the performance of human annotators. In case of the elements constituting references to literary works the performance of our models matches the humans. This is almost the case of the POI element as well.

4.2. Result analysis

The counts of detected elements closely correlate with the counts of annotations created by humans (compare first row of Table 2 with the third row of Table 1). The only exception is the implicit element which has been automatically recognized in only 158 cases whereas the humans found 483 instances of this element. This clearly suggests that our models struggled to recognize the implicit type of reference.

Overall the performance of the trained models was decent. It appears that in case of the l:author, l:title, l:other and POI the models almost matched human performance (compare the fourth row of Table 2 with the seventh row of Table 1). The models trained to recognize the c:id, c:type, c:date, and c:court constituents perform somewhat worse than human annotators.

It may be quite surprising to see the relatively low performance for elements such as c:date or c:court. Indeed the task of detecting dates or court mentions should not be that challenging. However, our models need to deal with a situation where we detect only certain dates and court mentions—only those that are part of references. Therefore the models may get confused by seeing mentions that appear to be of the relevant types but they are not. This problem could be mitigated in later stages where the constituents would be linked together to form references.

4.3. Grouping Constituents into References and Linking References to Content

Eventually we would like to use the automatically recognized constituents as building blocks for references. Grouping the constituents presents an interesting research problem in its own right. We already have the annotations that group the elements into individual references. A reference is essentially a set of a number of constituents. The detailed statistics on the references are reported in Table 3.

Finally we would like to connect each reference to a content element. The evaluation of the human annotator's effort is summarized in Table 4. The top two rows are the same types of measures as the ones used for references. The only difference is that the content is used as an additional constituent in all the four reference types.

	c:ref (expl)	c:ref (impl)	l:ref (expl)	l:ref (impl)
reference count	7570	1040	2497	293
average per doc	21.63	2.97	7.13	0.84
gold count	3753	429	1228	122
average gold per doc	10.72	1.23	3.51	0.35
strict agreement (inter)	45.96	20.66	33.23	22.58
overlap agreement (inter)	88.60	35.00	85.18	50.51

Table 3. Statistics of references in the data set

	c:ref (expl)	c:ref (impl)	l:ref (expl)	l:ref (impl)
strict agreement (inter)	42.18	12.42	30.65	19.49
overlap agreement (inter)	89.05	35.67	86.30	51.19

Table 4. Agreement on references between human annotators

5. Conclusions and Future Work

We presented early results from our ongoing effort to automatically detect references in Czech case-law. With regard to future work, our annotation task also involved marking polarity of references, which was not discussed in this paper. As such, it needs to undergo similar evaluation as other types of annotation. Another partial task for automation is the creation of whole references from lower level constituents. Moreover, successful statistical recognition is only a single step in our research. Ultimate goal is to allow for creation of citation network of the Czech top-tier court decisions and leverage this network to investigate the concept of 'importance' of court decisions and scholarly works.

6. Acknowledgements

The authors would like to thank annotators and editors – František Kasl, Adéla Kotková, Pavel Loutocký, Jakub Míšek, Daniela Procházková, Helena Pullmannová, Petr Semenišin, Nikola Šimková, Tamara Šejnová, Michal Vosinek, Lucie Zavadilová, and Jan Zibner. JH gratefully acknowledges the support from the Czech Science Foundation under grant no. GA17-20645S.

References

[1] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, GATE: an Architecture for Development of Robust HLT Applications. Proceedings of the 40th Annual ACL meeting, pp. 168-175.
[2] Vincent Kríž, Barbora Hladká, Jan Dědek and Martin Nečaský. Statistical Recognition of References in Czech Court Decisions. Proceedings of MICAI 2014, Part I, pp. 51–61.
[3] John Lafferty, Andrew McCallum, Fernando Pereira, and others. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning*, ICML, Vol. 1. 282–289.
[4] Naoaki Okazaki. *CRFsuite: a fast implementation of Conditional Random Fields*. (2007).
[5] Marc van Opijnen. Canonicalizing Complex Case Law Citations. JURIX 2010, pp. 97–106.
[6] Monica Palmirani, Raffaella Brighi and Matteo Massini. Automated Extraction of Normative References in Legal Texts. Proceedings of ICAIL 2003, pp. 105–106.
[7] Yannis Panagis and Urška Šadl. The Force of EU Case Law: A multidimensional Study of Case citations. Proceedings of JURIX 2015, pp. 71–80.
[8] Patent US6856988. Automated system and method for generating reasons that a court case is cited.
[9] Paul Zhang and Lavanya Koppaka. Semantics-Based Legal Citation Network. Proceedings of ICAIL 2007, pp. 123–130.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-183

Utilizing Vector Space Models for Identifying Legal Factors from Text

Mohammad H. Falakmasir [a] and Kevin D. Ashley [a,b]
[a] *Intelligent Systems Program, University of Pittsburgh*
[b] *School of Law, University of Pittsburgh*

Abstract.
Vector Space Models (VSMs) represent documents as points in a vector space derived from term frequencies in the corpus. This level of abstraction provides a flexible way to represent complex semantic concepts through vectors, matrices, and higher-order tensors. In this paper we utilize a number of VSMs on a corpus of judicial decisions in order to classify cases in terms of legal factors, stereotypical fact patterns that tend to strengthen or weaken a side's argument in a legal claim. We apply different VSMs to a corpus of trade secret misappropriation cases and compare their classification results. The experiment shows that simple binary VSMs work better than previously reported techniques but that more complex VSMs including dimensionality reduction techniques do not improve performance.

Keywords. Vector Space Models, Legal Analytics, Semantic extraction

1. Introduction

An important target of argument mining efforts in the legal field has been to extract factors from case texts. See [1, Chapter 10]. Legal factors are stereotypical patterns of fact that tend to strengthen or weaken a side's argument in a legal claim. [2, p.27].

Factors are particularly important in trade secret law. Information may qualify as a *trade secret* if it:

> is secret in the sense that it is not generally known among or readily accessible to people in the wider community that normally deal with the kind of information; has commercial value because it is secret; and has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret.[1]

Misappropriation consists of:

> acquisition of a trade secret of another by a person who knows or has reason to know that the trade secret was acquired by improper means; or disclosure or use of a trade secret of

[1]Trade Related Aspects of Intellectual Property Rights (TRIPS). Agreement on Undisclosed Information. Section 7: Protection of Undisclosed Information, Article 39.

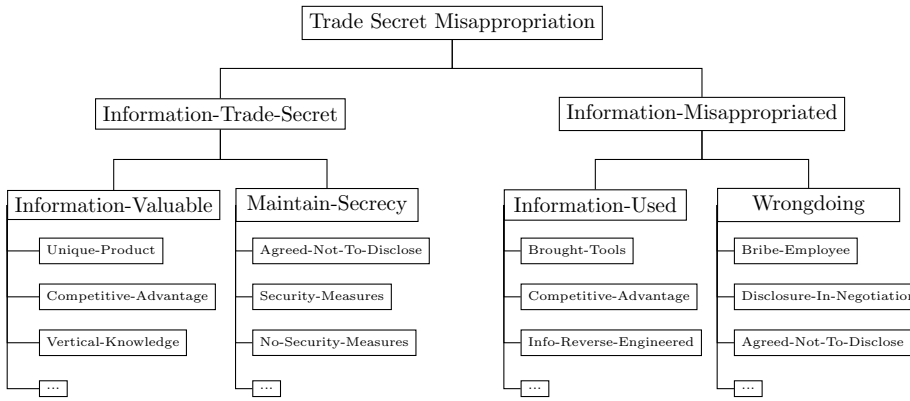


Figure 1. Example Factors from the Trade Secret Domain Model

> another without express or implied consent by a person who (i) used improper means to acquire knowledge of the trade secret ....[2]

Improper means include:

> theft, bribery, misrepresentation, breach or inducement of a breach of a duty to maintain secrecy, or espionage through electronic or other means [but not] reverse engineering, independent derivation, or any other lawful means of acquisition.[3]

A still influential secondary source of trade secret law introduced factors in an oft-cited set of guidelines for determining if information is a trade secret:

> An exact definition of a trade secret is not possible. Some factors to be considered in determining whether given information is one's trade secret are:
>
> 1. the extent to which the information is known outside of his business;
> 2. the extent to which it is known by employees and others involved in his business;
> 3. the extent of measures taken by him to guard the secrecy of the information;
> 4. the value of the information to him and to his competitors;
> 5. the amount of effort or money expended by him in developing the information;
> 6. the ease or difficulty with which the information could be properly acquired or duplicated by others.[4]

In the U.S. common law system, judges weigh the factors in a current case and explain their judgments by citing the statutes and guidelines and by making arguments based on past decisions or precedents. In modeling such case-base arguments, Ashley introduced dimensions to represent and elaborate the above factors into a set that ultimately comprised twenty-six factors, each favoring one side or the other [2]. For a complete list, see [3].

Ashley and Brüninghaus organized the claim requirements and factors into a domain model for the issue-based prediction system (IBP) [4]. Grabmair extended the model in the Value Judgment Formalism framework (VJAP) [5]. In this model (Figure 1) each factor is related to a high-level statutory requirement of a trade

[2] 18 U.S. Code 1839 - Definitions (5).
[3] 18 U.S. Code 1839 - Definitions (6)(B).
[4] Restatement (First) of Torts Section 757. Liability for Disclosure or Use of Another's Trade Secret. Comment b. Definition of trade secret states.

secret misappropriation claim. Each factor weighs in favor of one side or other. For example, [F6:Security-Measures] favors the plaintiff trade secret holder and indicates that it applied active measures to limit access and distribution of the information that is the property of interest. [F24:Info-Obtainable-Elsewhere] implies that the confidential information could be obtained from publicly available sources and favors the defendant, the alleged misappropriator.

For purposes of modeling, a conclusion that a factor applies in a case is based on classifying at least one sentence in the text as an instance of the factor. For example, the following sentences justify the conclusion that the associated factors apply in the *Mason* case, a trade secret dispute concerning the recipe for a drink, Lynchburg Lemonade (See [1, Figure 11.8]:)

- **F6:Security-Measures (pro-plaintiff):** He testified that he told only a few of his employees--the bartenders--the recipe. He stated that each one was specifically instructed not to tell anyone the recipe. To prevent customers from learning the recipe, the beverage was mixed in the "back" of the restaurant and lounge.
- **F15:Unique-Product (pro-plaintiff):** It appears that one could not order a Lynchburg Lemonade in any establishment other than that of the plaintiff.
- **F16:Info-Reverse-Engineerable (pro-defendant):** At least one witness testified that he could duplicate the recipe after tasting a Lynchburg Lemonade.
- **F21:Knew-Info-Confidential (pro-plaintiff):** On cross-examination Randle agreed that he had been under the impression that Mason's recipe for Lynchburg Lemonade was a secret formula.

Our main research goal is to improve the performance of automatically classifying the texts of trade secrets misappropriation cases by their applicable factors. As an initial step we asked how well vector space models (VSMs) can identify factors in the case texts (see Section 2). Using the domain model of Figure 1, once factors are identified, one could also identify the legal issues litigated in the case.

In our study, eight different VSMs plus variations learn different representations of the case texts in our corpus. Four of the VSMs are based on relatively simple binary or TF-IDF representations (see section 3). The other four employ dimensionality reduction techniques to represent case texts. We compared the representations learned by the different VSMs in terms of their results on classifying a subset of a gold standard corpus of 172 cases tagged by legal experts as to applicable factors. Initially, we hypothesized that the dimensionality reduction techniques would lead to learning VSMs that were more expressive of the underlying legal factors. Based on the results reported below, we can reject that hypothesis. Nevertheless, all of the VSMs outperformed previously-reported results in classifying case texts by legal factors.

2. Background on Vector Space Models

Statistical studies of semantics represent meaning as a probability distribution over a set of latent dimensions using the bag-of-words hypothesis or the distributional hypothesis [18]. The key idea is that if units of text have similar vectors in a term frequency matrix, they tend to have similar meaning.

Based on the bag-of-words hypothesis, word frequencies in a document indicate the relevance of the document to a search query. Given a large corpus, one

can form a term-document matrix where the rows correspond to terms and the columns correspond to the frequencies of words in each document. Most of the elements of the term-document matrix are zero since most documents use only a small fraction of the whole vocabulary. The term-document matrix provides a very broad notion of meaning that is suitable for document retrieval. However, it only supports a coarse-grained measure of topical similarity [18].

Based on the distributional hypothesis, words that appear together in the same context tend to have similar meaning. The context could be a sentence, or perhaps even a fixed window of words. In general, shorter windows tend to capture syntactic features while longer windows tend to capture more semantic relations. The distributional hypothesis is the main inspiration of the recent neural network-based models for learning word vectors (word embeddings a.k.a word2vec) [18].

Vector space models of semantics represent meaning as a coordinate in a high-dimensional "semantic space". Vector representations are a common way to compute semantic similarity between arbitrary spans of text. Each context vector is a point in $|V|$-dimensional space. $|V|$, the length of the vector, is generally the size of the vocabulary. Quite often, raw term frequencies (TFs) are not the best measure of semantic similarity because word frequencies follow a skewed distribution according to Zipf's Law. An alternative measure of similarity between documents is TF-IDF. The TFs are often weighted by the inverse document frequency (IDF) to give a higher weight to rare words that occur only in a few documents [18]. The nature of the vectorized representation allows documents to be compared in terms of semantic similarity using any of the standard similarity or distance measures available from linear algebra (e.g., cosine similarity or Euclidean distance).

One can also apply various dimensionality reduction techniques, such as singular value decomposition (SVD), non-negative matrix factorization (NMF), and Latent Dirichlet Allocation (LDA) [18]. These methods can essentially be thought of as a way to cluster words along a small number of *latent* semantic dimensions that are automatically learned from a low-rank approximation of the term-document matrix. In fact, Latent Semantic Analysis (LSA) is a low-rank approximation of the term-document matrix using SVD, and both LDA and NMF has been successfully applied in the literature for topic modeling [18].

3. Data and Methods

For this study, we compiled a corpus of trade secret misappropriation cases by scraping the texts of 1,600 federal and state opinions retrieved from the CourtListener website[5] that contain references to two particular sources of legal rules: (1) the Restatement of Torts section 757, comment b (1939) ("RT757") and (2) Uniform Trade Secrets Act (1985) ("UTSA"). We also used a gold-standard corpus of 172 cases from the HYPO, CATO, SMILE, and VJAP programs (VJAP corpus) whose sentences legal experts labeled according to the 26 trade secret factors. Table 1 provides summary statistics of these corpora. The totals correct for the fact that some of the cases cite both references.

[5] https://www.courtlistener.com/

Table 1. Summary statistics of the available corpora.

Corpora	# Cases	# Sentences	# Terms	# Verbs
Restatement of Torts 757	509	108,186	36,454	26,630
Uniform Trade Secret Act	1,213	226,556	52,232	36,973
VJAP (based on HYPO, CATO, IBP)	179	26,296	19,327	13,884
Total (Unique Cases)	1,600	334,742	62,472	44,559

The performance of machine learning methods depends heavily on the choice of data representation (or features) to which they are applied. Domain knowledge can be an important resource for designing effective text representations. Feature-engineering is labor-intensive, however, and domain models evolve over time. Ideally, a representation would capture the underlying distribution of the data and automatically account for the evolution of these abstractions. Our goal in this project is to see how far one can go without feature-engineering. At the same time, we remain open to applying techniques to efficiently incorporate domain knowledge where feasible, a task for future work.

We designed our experiments as a four-step pipeline. The first step (preprocessing) includes tokenization, part-of-speech tagging, and extracting main verbs of the sentence. In the second step (vectorization) we learn multiple vector representations for the opinions in the case base. Then, we form the term-document matrix for the corpus based on the bag-of-words hypothesis.

Our binary VSM models, bag-of-words (BOW) and bag-of-verbs (BOV) represent each document using one-hot encoding, that is, with one Boolean column for each category. Since many factors correspond to parties' actions, we created the bag-of-verbs version of the term-document matrix [19]. We use a modified form of the verb by concatenating the immediate conjunct of the verb according to the dependency parse results. For example in this form, we have separate tokens for the verb "disclose" including, disclosed, not_disclosed, not_to_disclose, have_disclosed, etc. This way of representing verbs is different from the forms used in the topic modeling literature that often uses the stemmed version of the verbs ("disclos" for all of the above forms) and removes the conjuncts as stop words. The main reason is that in the legal context in general and in considering the factors in particular, the verb tense and the modals play a pivotal role in the fact finding process of the decision maker and should not be mapped into the same dimension (considered as the same "token") in the feature space.

The next two VSMs, TF-IDF (Terms) and TF-IDF (Verbs) are standard TF-IDF transformations on the term-document or verb-document matrix. We use document frequencies and raw counts as a filter to remove case-specific information. Since our evaluation set (VJAP corpus) only contains 172 carefully selected opinions (not a random sample) we use the larger corpus of 1,600 scraped opinions to calculate the counts and apply the TF-IDF weights.

In the third step (transformation) we apply four widely used VSM models to reduce the dimensionality of our representations and infer latent dimensions. Latent Semantic Analysis (LSA) and Non-negative matrix factorization (NMF) are two of the numerical approaches for transforming documents into a semantic vector spaces. Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) are probabilistic alternatives for inferring latent dimensions.

In the fourth step (evaluation) we used the VSM representations of the documents from the VJAP corpus in a supervised classification framework to predict the factors and to investigate the association of each VSM with the targeted labels, the legal factors. We evaluate all of the representations (i.e., BOW, BOV, TF-IDF (Terms), TF-IDF (Verbs), LSA, NMF, LDA, and HDP) by comparing their resulting classification results.

4. Experiments

A key aspect of creating an expressive VSM model is choosing the right size of representations K for different applications (datasets). This is provided as a parameter of the model (similar to parameter K of k-means clustering). Assuming that there is a finite number of legal factors one can assume that each case is a point in a 26-dimensional space. The classifier is finding a *surface* that has only points with positive labels on one side and points with negative labels on the other side. There are numerical methods for identifying the right K; however, in this study we experiment with different values for K to investigate the effect of the size of representation for the task at hand.

We start with a term-document matrix ($X_{m\times n}$) with real-valued, non-negative entries (TF-IDF weights). Among the various ways of learning document representations, this paper focuses on low-rank approximation of the term-document matrix in the form of:

$$X_{m\times n} = W_{m\times r} \times H_{r\times n} \qquad r < \min(m, n) \tag{1}$$

The term-document matrix X is factorized into two smaller matrices, W as a document archetype that encapsulates the intensity (weight) of each term in the feature space and H that represents the projection of each document into that feature space. m is the number of terms in our corpus, n is the number of documents, and r is the dimension of our representation (feature space).

For the bag-of-words (BOW) and bag-of-verbs (BOV) VSMs, the vectorized representation of each case is a $|V|$ dimensional one-hot vector ($[0, 1, 1, 0, 0]_{|V|}$) that is created by considering terms and verbs. The TF-IDF (Terms) and (Verbs) VSMs use a standard TF-IDF weighting with n-grams (n=1, 2, 3). We filter out terms that appeared in more than 90% of the documents or fewer than 5 times throughout the corpus. We also report the results of four widely used VSM models (LSA, NMF, LDA, HDP) and experimented with different Ks to find the optimum number of dimensions. The HDP model is a non-parametric method and does not require the number of dimensions to be specified in advance. All of the experimental models can be considered as a relaxed form of k-means clustering, with columns of the W representing the cluster centroids and rows of the H indicating cluster membership (weights) for each document. As a result, the output of our VSMs are r dimensional vectors for each document.

For evaluation, we used a Support Vector Machine (SVM) with a linear kernel in a multi-label (One-vs-Rest) classification framework. We train a binary SVM classifier for each factor without performing any parameter optimization. Although one can tune the C parameter of the SVM classifier to increase the recall at the expense of lower precision [16], we decided to use F1 as our evaluation

Table 2. Experimental Results

#	VSM / Results	Precision (mi/ma)	Recall (mi/ma)	F1 (mi/macro)	#Features
1	BOW	0.86/0.80	0.50/0.49	0.63/0.58	20,001
2	BOV	0.90/0.80	0.49/0.48	0.63/0.58	13,649
3	TF-IDF (Terms)	0.80/0.75	0.49/0.48	0.61/0.56	406,641
4	TF-IDF (Verbs)	0.89/0.81	0.52/0.49	**0.65/0.59**	46,373
5	LSA (20)	0.38/0.32	0.63/0.61	0.47/0.40	20
6	LSA (50)	0.52/0.46	0.64/0.62	0.58/0.50	50
7	LSA (100)	0.63/0.60	0.58/0.56	0.61/0.56	100
8	LSA (200)	0.73/0.75	0.55/0.53	0.62/0.59	200
9	LSA (400)	0.88/0.79	0.52/0.50	**0.65/0.59**	400
10	NMF (50)	0.26/0.35	0.50/0.52	0.34/0.33	50
11	NMF (100)	0.26/0.31	0.53/0.51	0.35/0.30	100
12	LDA (20)	0.27/0.23	0.56/0.57	0.37/0.31	20
13	LDA (50)	0.30/0.25	0.52/0.48	0.38/0.31	50
14	LDA (100)	0.38/0.35	0.58/0.56	0.46/0.41	100
15	LDA (200)	0.48/0.41	0.58/0.55	0.52/0.46	200
16	LDA (400)	0.51/0.43	0.59/0.55	0.55/0.45	400
17	HDP	0.45/0.41	0.51/0.50	0.48/0.43	150

metric which is a harmonic mean of precision and recall. We used 70% percent of the documents in our corpus for training the classifiers and 30% of the documents as a hold-out test set in a stratified fashion. We thus ensure the distribution of the target labels is roughly the same in our training and test sets.

5. Results and Discussion

Table 2 shows the results. We report precision, recall, and F1 scores both on the micro and macro level, but our main evaluation metric is macro F1.

TF-IDF (Verbs), the TF-IDF model that used n-grams of the verbs, and LSA (400) did best. On the positive side, this performance (and that of all the VSMs tested) is better than that of previously reported efforts (see Section 6).

On the other hand, the best-performing VSMs did only slightly better than the BOW or BOV models. These binary VSM models outperformed most of the VSM models with more complex, dimensionality-reducing representations. Since we expected that more complex VSMs might better reflect the legal factors and be better able to minimize feature engineering effort, this result was disappointing.

One way to explain these results is that our gold-standard corpus provides labels at the document level while trade secret factors are usually discussed on a sentence level, a problem also pointed out in [15]. Moreover, the labels are annotated mainly to study the interaction of factors in the trade secret domain, and there are some false negatives due to cases where the factor was mentioned but not applied in the decision (e.g., the factor may have been discussed in a description of a case cited for other reasons.) In an ideal scenario, the document representation should be able to filter-out noise and irrelevant case-specific information from the raw text files and aggregate information that discusses the factors actually applied and the issues actually decided. This may require the identification of sentence role types such as court's findings of fact. See [1, Chapter 11].

Table 3 shows the results of a best-performing model, TF-IDF (Verbs). Some factors may not have enough examples from which to learn. There also are excep-

Table 3. Results of a Best Model (TF-IDF Verbs)

Factor	P	R	F1	# ts	# tr	Top 5 Features (n-grams)
Security-Measures	0.76	0.9	0.83	29	62	hinged, submitted, think, snap, sold,
Info-Independently-Generated	0.89	0.53	0.67	15	43	tying, not_to_compete, snap, provided, find,
Disclosure-In-Negotiations	0.82	0.64	0.72	14	39	testified, not_to_compete, using, to_develop, taken,
Agreed-Not-To-Disclose	0.78	0.44	0.56	16	38	to_sell, sold, prevailing, hinged, design,
Brought-Tools	1	0.89	0.94	9	27	tying, disclosed, disclosing, hinged, affirm,
Restricted-Materials-Used	1	0.33	0.5	15	27	not_to_compete, erred, sold, shows, prevailing,
Info-Known-To-Competitors	1	0.44	0.62	18	24	shows, said, making, had, found,
Identical-Products	1	0.71	0.83	7	23	manufacturing, testified, developed, said, argues,
Disclosure-In-Public-Forum	1	0.33	0.5	12	21	said, contend, sitting save, sitting save read, given,
Unique-Product	1	0.3	0.46	10	21	found, not_to_compete, said, became, appropriated,
Secrets-Disclosed-Outsiders	1	0.38	0.55	8	17	manufacturing, denied, disclosed, were, claimed,
Outsider-Disclosures-Restricted	1	0.5	0.67	4	15	find, denied, sitting, submitted, desired,
No-Security-Measures	0	0	0	10	14	contained, said, think, found, selling,
Bribe-Employee	1	0.6	0.75	5	13	contend, enjoined, affirmed, implied, think,
Deception	0	0	0	3	13	argues, were, developed, acquired, sitting,
Agreement-Not-Specific	1	0.33	0.5	3	12	contained, not_to_compete, concluded, contend, conclude,
Vertical-Knowledge	1	0.6	0.75	5	11	said, think, existed, employed, continued,
Competitive-Advantage	0.75	0.6	0.67	5	10	erred, cited, tying, disclosing, affirmed,
Waiver-Of-Confidentiality	1	0.25	0.4	4	9	disclosed, found, denied, contained, said,
Employee-Sole-Developer	1	1	1	1	8	found, find, held, argues, affirmed,
Info-Reverse-Engineered	1	0.6	0.75	5	7	testified, not_to_compete, reverse, contend, using,
Noncompetition-Agreement	1	0.67	0.8	3	6	denied, tied, think, concerning, referred,
Info-Obtainable-Elsewhere	0	0	0	1	6	found, to_be, said, testified, using,
Invasive-Techniques	1	1	1	1	4	denied, were, developed, found, manufacturing,
Knew-Info-Confidential	1	0.75	0.86	4	4	were, concerning, erred, claimed, known,
Info-Reverse-Engineerable	0	0	0	4	1	testified, denied, found, contained, using,

tions like [Employee-Sole-Developer] or [Invasive-Techniques] that resulted in perfect classification despite the lack of training data and factors like [No-Security-Measures] that resulted in F1 score of 0.0 despite having 24 examples cases in the gold-standard corpus. We could explain this observation based on the fact that we have another factor [Security-Measures] which is closely related to the [No-Security-Measure] factor and might have caused some ambiguity for the classification. One could update the domain model based on this observation and merge these two factors into a single binary factor that takes values True or False.

The five most predictive features for each factor in Table 3 indicate that the SVM classifier has learned some promising features. [Info-Independently-Generated] and [Disclosure-In-Negotiations] each have the verb "not_to_compete" with F1-scores of 0.67 and 0.72 respectively. These results shows some potential for applying VSMs in the legal domain with minimal domain modeling.

6. Related Work

As noted, all of the VSM models outperformed the results reported by Ashley and Brüninghaus [4] with respect to the macro F1. In project SMILE, the researchers tested three representation schemes trying to predict the factors from the IBP corpus (of which VJAP's case base is a subset) [4] [14]. The first representation (BOW) was a bag-of-words representation similar to that of our bag-of-words VSM. In the second representation (RR), they replaced the parties and product names with their roles in the case. The third representation (ProPs) utilized the dependency parse results and converted each sentence within the case into (sub-

ject, verb), (verb, object), (verb, prepositional phrase), and (verb, adjective) tuples. They also performed additional processing to the negated verbs and passive verb forms within each sentence. However, the results were suboptimal (reported average F1=0.21) mainly due to the large dimensionality of the *bag-of-words* space and the lack of training data for each factor. Wyner and Peters [15] tried to solve this problem by starting from the description of the factors and using WordNet[6] expansions and expert knowledge to generate *factoroids*, plausibly semantic terms that are related to each factor. They used factoroids to generate rules as a part of GATE system[7] to annotate cases with respect to factors and pointed out the utility of creating a gold-standard corpus for machine learning.

In e-discovery, unsupervised learning enables exploratory clustering of documents and selecting seed sets for supervised learning. For example, the Categorix system clusters documents for review using PLSA, a probabilistic alternative to LSA as we used [6]. In earlier work, Uyttendaele, et al. applied an unsupervised, non-hierarchical clustering method and a TF-IDF vector space model like our TF-IDF (Terms) VSM to group paragraphs in court opinions thematically for purposes of summarization [7]. Schweighofer and Merkl applied self-organizing maps, a kind of unsupervised neural network, to explore and cluster documents in a corpus of European legal texts concerning public enterprises [8].

Lu, et al. clustered and segmented legal documents by topic in a huge corpus including judicial opinions and statutes [9]. The clustering process, however, used metadata unavailable to us including document citations, user behavior data, and topical classifications, which do not appear to capture topical information as detailed as trade secret factors. Winkels, et al. applied unsupervised learning to identify natural clusters of case citations of statutes, (as opposed to clusters of cases themselves as we do) for eventual use in a legal recommender context [10].

More recently, Panagis, et al. applied non-negative matrix factorization to a large set of judgments from the EU Court of Justice and European Court of Human Rights and selected clusters using topic coherence via word2vec to study topic drift over time [11]. Landthaler, et al. employed word embeddings (word2vec) in extracting similar obligations from the text of an EU Data Protection Directive 94/46/EC (EU-DPD) and similar provisions from a collection of German rental contracts [12]. We used NMF but did not employ word embeddings and leave it for future work as a potential substitute for WordNet expansion. Most recently, McCarty has called for an unsupervised approach to learning legal semantics in a corpus of unannotated cases to generate structured case notes [13].

7. Conclusion and Future Work

We used Vector Space Models to identify legal factors in in trade secret misappropriation cases. Factors, complex categories that capture a claim's substantive strengths and weaknesses, are intermediaries between statutory legal elements and cases' particular facts. Our results show that with simple heuristics and off-the-shelf components, one can detect some signal (i.e., features) for classifying

[6]https://wordnet.princeton.edu/
[7]https://gate.ac.uk/

factors in case texts. Our VSMs performed better than a previously published attempt at learning to identify factors in cases. On the other hand, our simplest VSMs outperformed most of the more complex ones, suggesting that dimensionality reduction did not add much if anything to classification performance.

We will study the *latent* dimensions learned by LSA, LDA, NMF, or HDP, to find mappings between what the model learns and legal factors. For example, one of the LSA model's latent dimensions contains the following verbs ordered by frequency: was used, was acquired, and to make, produce, obtain, manufacture, solicit, determine, establish, develop, show, prevent, design, gain, compete, treat, and enjoin. This latent dimension is related to the [Information-Used] branch of the domain model (Figure 1). One may learn a frequency, top-n threshold, or heuristic to identify factors with a finer granularity under this branch.

Such methods may pre-process case texts with factor-related information, so that human reviewers can confirm factor classifications more efficiently.

References

[1] Ashley, K. Artificial Intelligence and Legal Analytics. Cambridge University Press, 2017.
[2] Ashley K. Modeling Legal Argument: reasoning with cases and hypotheticals. The MIT Press, Cambridge, Massachusetts; 1990.
[3] Aleven V. Teaching case-based argumentation through a model and examples Ph.D. thesis, University of Pittsburgh, Pittsburgh, PA, USA.
[4] Ashley K, Brüninghaus S. Computer models for legal prediction. Jurimetrics 2006:309-352.
[5] Grabmair M. Modeling Purposive Legal Argumentation and Case Outcome Prediction using Argument Schemes in the Value Judgment Formalism: U. Pittsburgh; 2016.
[6] Privault C, ONeill J, Ciriza V, Renders J-M. A new tangible user interface for machine learning document review. Artificial Intelligence and Law 2010;18(4):459-479.
[7] Uyttendaele C, Moens M-F, Dumortier J. Salomon: automatic abstracting of legal cases for effective access to court decisions. Artificial Intelligence and Law 1998;6(1):59-79.
[8] Schweighofer E, Merkl D. A learning technique for legal document analysis. ICAIL 1999. ACM Press. p 156-163.
[9] Lu Q, Conrad J, Al-Kofahi K, Keenan W. Legal document clustering with built-in topic segmentation. 2011. 20th ACM Int'l Conf. Info. and Knowledge Management. p 383-392.
[10] Winkels R, Boer A, Vredebregt B, van SOMEREN A. Towards a Legal Recommender System. 2014. JURIX. 271:169-178.
[11] Panagis Y, Christensen ML, Sadl U. On Top of Topics: Leveraging Topic Modeling to Study the Dynamic Case-Law of International Courts. JURIX 2016. p 161-166.
[12] Landthaler J, Waltl B, Holl P, Matthes F. Extending Full Text Search for Legal Document Collections Using Word Embeddings. JURIX 2016. p 73-82.
[13] McCarty LT. Discussion Paper: On Semi-Supervised Learning of Legal Semantics. 2017.
[14] Ashley KD, Brüninghaus S. Automatically classifying case texts and predicting outcomes. Artificial Intelligence and Law 2009;17(2):125-165.
[15] Wyner AZ, Peters W. Lexical Semantics and Expert Legal Knowledge towards the Identification of Legal Case Factors. JURIX 2010. p 127-136.
[16] Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. Journal of machine learning research. 2008;9(Aug):1871-4.
[17] Moschovakis YN. Sense and denotation as algorithm and value. Lecture notes in logic. 1994;2:210-49.
[18] Jurafsky D, Martin JH. Vector Semantics. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed draft chapter 15-16). 2017.
[19] Wijaya DT. VerbKB: A Knowledge Base of Verbs for Natural Language Understanding (Doctoral dissertation, Carnegie Mellon University).

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-193

Concept Recognition in European and National Law

Rohan NANDA [a,b,1], and Giovanni SIRAGUSA [a] and Luigi DI CARO [a] and Martin THEOBALD [b] and Guido BOELLA [a] and Livio ROBALDO [b] and Francesco COSTAMAGNA [a]

[a] *University of Turin, Italy*
[b] *University of Luxembourg*

Abstract. This paper presents a concept recognition system for European and national legislation. Current named entity recognition (NER) systems do not focus on identifying concepts which are essential for interpretation and harmonization of European and national law. We utilized the IATE (Inter-Active Terminology for Europe) vocabulary, a state-of-the-art named entity recognition system and Wikipedia to generate an annotated corpus for concept recognition. We applied conditional random fields (CRF) to identify concepts on a corpus of European directives and Statutory Instruments (SIs) of the United Kingdom. The CRF-based concept recognition system achieved an F1 score of 0.71 over the combined corpus of directives and SIs. Our results indicate the usability of a CRF-based learning system over dictionary tagging and state-of-the-art methods.

Keywords. Concept Recognition, European Law, Information Retrieval

1. Introduction

With the increasing volume of European and national legislation available online, the identification of domain concepts in legal texts is very important for the development of legal information retrieval systems. The identification of domain concepts provides a deeper insight into the interpretation and understanding of texts. The recognition of concepts in legal texts would also be useful for the harmonization and integration of European and national law. Research in this domain has mainly focused on identification of named entities like person, organization and location names. However, European and national legislation contains very few instances of named entities. They primarily comprise legal and domain-specific jargon which can be represented by concepts.

In this paper, we develop a system for concept recognition in European directives and national law (statutory instruments of the United Kingdom). The concept recognition system was used for automatically identifying concepts in a corpus of 2884 directives and 2884 SIs. We generated an annotated corpus using a semi-supervised approach to

[1]Corresponding Author

save human effort and time for evaluation of our system. Further, we also generated a mapping to link similar terms in directives and SIs under the same concept.

The rest of the paper is organized as follows. In the next section, we discuss the related work. Section 3 describes the concept recognition system. Section 4 discusses the results and analysis. The paper concludes in Section 5.

2. Related Work

Related work is mainly focused in the domain of named entity recognition (NER) systems. In [1], the authors developed a legal named entity recognizer and linker by aligning YAGO[2] (WordNet-and Wikipedia-based ontology) and the LKIF ontology. The alignment was carried out manually by mapping a concept node in LKIF to its equivalent in YAGO. They utilized different models like support vector machines (SVM), Stanford Named Entity Recognizer (NER) [4], and neural networks and evaluated the system on a small sample of judgements from the European Court of Human Rights (ECHR). Their results indicate that the LKIF level of generalization is not suitable for named entity recognition and classification as their system was unable to distinguish between the classes defined in LKIF. However, their NER system achieved a better performance while distinguishing YAGO classes. The authors in [3] developed a named entity recognition and classification system to recognize entities like judges, attorneys, companies, courts and jurisdictions in US case law, depositions, pleadings and other trial documents. They utilized dictionary lookup, contextual pattern rules and statistical models for identifying named entities. The NER system was trained using a SVM classifier and evaluated on manually and automatically acquired training datasets of case law. The authors in [2] developed a NER system using AdaBoost. The system uses a window, along with a set of features (part-of-speech tags and dictionary of words) to capture the local context of a word. Current NER systems are based on conditional random fields (CRF) [5], which allow to train a unique model for the classification and recognition of named entities. In [4], the authors developed a CRF which used Gibbs sampling instead of the standard Viterbi algorithm. They demonstrated that the use of Gibbs sampling allowed the system to distinguish between mentions of organization or person on the basis of context, thus enforcing label consistency.

3. Concept Recognition System

In this section, we describe the concept recognition system for European and national law. In the legal domain, concepts are generally represented using ontologies or vocabularies. Previous NER systems (based on the concepts represented in the LKIF ontology) demonstrated that the LKIF level of generalization was not suitable [1]. This is because NER systems could not clearly distinguish between the classes defined in LKIF. Therefore, in this paper we investigate the use of vocabularies for developing our concept recognition system. We utilize Inter-Active Terminology for Europe [3] (IATE), which is the EU's inter-institutional terminology database. IATE consists of 1.3 million entries in

[2]http://www.yago-knowledge.org/

[3]http://iate.europa.eu

English. Every entry (concept) in IATE is mapped to a subject domain. We filtered out some irrelevant entries in IATE (stopwords and concepts mapped to 'NO DOMAIN').

We utilized a corpus of 2884 directives and 2884 statutory instruments for our experiments. Since training data was not available, we utilized a semi-supervised approach to generate an annotated corpus. The development of NER or concept recognition systems require a large amount of manually annotated datasets, which are expensive to obtain. We manually annotated a few documents with IATE subject domains. Then we developed a dictionary lookup program to tag terms (both words and phrases) in the text with IATE subject domains. Each term in the text was compared to entries in the IATE vocabulary and matching terms were tagged with the relevant subject domains. Table5 shows some examples of these terms and subject domains. We also used spaCy[4], a state-of-the-art NER system to annotate time, date and monetary units. We filtered out irrelevant candidate entities by using Dexter [7], a Wikipedia entity linker.[5] Then, we annotated all the documents in the corpus. After generating the annotated corpus for both directives and SIs we divided each dataset into an 80% training (2307 documents) and a 20% test set (577 documents) to build the concept recognition system. The combined corpus comprised 80% training set (2307 directives + 2307 SIs) and 20% test set (577 directives + 577 SIs). Table 1 shows the number of documents, tokens and vocabulary size for both the directive and SI datasets, respectively. We observe that SIs have a much larger vocabulary than directives. We utilized conditional random fields (CRFs) to build our concept

Table 1. Number of documents, number of tokens and the vocabulary size (|V|) for directives (left) and SIs (right). We computed $|V_{total}|$ as $|V_{train}| + |V_{test}| - |V_{train} \cap V_{test}|$

Dataset	# docs	# tokens	|V|
Train	2,307	4,646,286	24,522
Test	577	1,226,338	14,127
Total	2,884	5,872,624	38,649

Dataset	# docs	# tokens	|V|
Train	2,307	4,189,157	83,172
Test	577	1,096,246	33,757
Total	2,884	5,285,403	116,929

recognition system as they have been known to work well in tasks which require labeling sequence data (especially natural language text). They are discriminative probabilistic models where each observation is a token from a sentence and the corresponding label (tag of subject domain or entity) represents the state sequence. We utilize the following features for our CRF model: word suffix, word identity (whether a word represents a subject domain/named-entity or not), word shape (captialized, lowercase or numeric) and part-of-speech (POS) tags. We used the limited-memory BFGS training algorithm with L1+L2 regularization.

4. Results and Analysis

In this section, we present the results of our system. Table 2 reports the F1 score of our CRF-based concept recognition model for each subject domain and entity class. We observe that all IATE subject domains are clearly distinguished due to the achievement of a reasonable F1 score for each domain for each corpus. The lower F1 score of domain 'INTERNATIONAL ORGANISATIONS' and named entities like 'QUANTITY', 'MONEY' and 'ORDINAL' is explained by a smaller number of tagged tokens, resulting

[4]https://spacy.io/

[5]Wikipedia Entity Linkers find named entities in the text that can be linked to a Wikipedia page.

in only a few training instances The other subject domains had sufficient training data and therefore were classified with a higher F1 score. These results also indicate that European and national legislation consist of very few named entities and are therefore more suited for concept recognition.

Table 2. Results (F1 score) for concept recognition for each class by the CRF-based concept recognition system

Tag name	Directives	SIs	Directives + SIs
IATE Subject Domains			
FINANCE	0.68	0.62	0.62
POLITICS	0.70	0.74	0.71
ENVIRONMENT	0.68	0.41	0.66
EDUCATION AND COMMUNICATIONS	0.68	0.72	0.71
LAW	0.92	0.81	0.89
INTERNATIONAL ORGANISATIONS	0.52	0.14	0.32
EMPLOYMENT AND WORKING CONDITIONS	0.70	0.68	0.70
AGRI-FOODSTUFFS	0.75	0.73	0.68
INDUSTRY	0.67	0.45	0.60
PRODUCTION TECHNOLOGY AND RESEARCH	0.69	0.67	0.69
BUSINESS AND COMPETITION	0.78	0.77	0.77
ENERGY	0.81	0.50	0.74
TRANSPORT	0.59	0.60	0.58
EUROPEAN UNION	0.79	0.77	0.76
AGRICULTURE FORESTRY AND FISHERIES	0.70	0.58	0.64
SOCIAL QUESTIONS	0.68	0.65	0.66
ECONOMICS	0.66	0.57	0.68
GEOGRAPHY	0.52	0.76	0.75
INTERNATIONAL RELATIONS	0.70	0.59	0.59
SCIENCE	0.60	0.48	0.59
TRADE	0.77	0.66	0.76
spaCy Named Entities			
QUANTITY	0.00	0.00	0.00
MONEY	0.00	0.00	0.00
ORDINAL	0.00	0.00	0.00
TIME	0.62	0.00	0.60
DATE	0.00	0.19	0.47

Table 3. Results of concept recognition with CRF model and comparison with a baseline ("Most frequent class") and the Stanford NER model

Corpus	System	Precision	Recall	F1 score
Directive Corpus	Most frequent class	0.74	0.53	0.61
	CRF	0.80	0.71	0.75
	Stanford NER	0.80	0.71	0.75
SIs Corpus	Most frequent class	0.61	0.40	0.48
	CRF	0.73	0.61	0.66
	Stanford NER	0.68	0.53	0.59
Combined Corpus (Directives + SIs)	Most frequent class	0.66	0.47	0.54
	CRF	0.76	0.68	0.71
	Stanford NER	** (did not finish training)	**	**

The average F1 scores of our CRF-based concept recognition system for directive, SI and combined corpus were 0.75, 0.66 and 0.71 respectively (Table 3). We also compare the performance of the CRF with a baseline method (the "Most frequent class" model). We observe that the CRF outperforms the baseline model. This is because the baseline model does not take into account the context information for a particular token while assigning it to a class. We also compared the CRF with Stanford NER for both the directive corpus and the SIs corpus. The CRF model had similar performance to the Stanford NER

in the directive corpus. However, it outperformed the Stanford NER in the SIs corpus by achieving a higher F1 score. For the combined corpus,the Stanford NER was still in training and we could not record the results in time (Stanford NER takes several days for training, perhaps due to use of long n-gram sequences). These runs are indicated by ** in Table 3. Our CRF system did not include n-gram features.

One drawback of using dictionary tagging to annotate a corpus is that some terms are missed and not tagged due to inconsistent rules to accommodate different phrases and tokenization errors. In the IATE dictionary, an entry, e.g., 'integrated energy performance', is linked to a subject domain, e.g., 'INDUSTRY'. Table 4 presents an example sentence with tagged labels of the IATE dictionary and predicted CRF labels. The CRF classifies both 'energy' and 'performance' to the 'INDUSTRY' subject domain, whereas the dictionary missed them. This is because the dictionary lookup utilizes state-of-the-art tokenizers which may not be 100% accurate and may lead to an incorrect tokenization, thus resulting in a mismatch. The CRF on the other hand, had some training instances from which it learns that the terms 'energy' and 'performance' are related to 'INDUSTRY'. Thus it was able to correctly classify them. Therefore, training a CRF model is advantageous also on automatically annotated corpora because it can improve the tagging of the dictionary by learning these semantic relations between terms and subject domains. Thus, it can be used to improve the quality of annotations and develop a better gold standard for further work.

Table 4. Comparison of CRF output with the dictionary tagging

	CRF predicted labels	Dictionary
calculation	O	O
of	O	O
the	O	O
integrated	O	O
energy	**INDUSTRY**	**O**
performance	**INDUSTRY**	**O**
of	O	O
buildings	O	O

In order to utilize the concept recognition system, it is important to align similar terms across European and national law. This semantic alignment of terms is highly useful for legal professionals to understand the differences in terminologies at the European and national level. The concept recognition system generates a large collection of terms under each subject domain from both directives and statutory instruments. We divided the terms under each subject domain into two lists: directive terms and SI terms. We computed the set difference of these two lists to obtain a list of terms present in the directives but not in the SIs. Similarly, we also obtained a list of terms present in SIs but not in the directives. We then computed text similarity (using Levenshtein distance) to find the most semantically similar term in the SIs (but not present in the directives) for a particular term in the directive. Table 5 shows a few examples of such terms. In future work, we intend to use the mapping of such terms to extend our text similarity system of detecting also transposing provisions for EU directives [6].

5. Conclusion

In this paper, we developed and evaluated a CRF-based concept recognition system for European and national law. We generated a labeled corpus of directives and statutory

Table 5. Aligned terms from European and national law

Subject Domain	Aligned terms (*Directive→SIs*)
EMPLOYMENT AND WORKING CONDITIONS	*professional qualification→vocational qualification* *seniority→job security* *occupational disease→industrial disease*
FINANCE	*life assurance→endowment assurance* *financial institution→financial administration* *dividend→tax on dividends*

instruments with subject domains of the IATE vocabulary, Wikipedia and a state-of-the-art named entity recognition system. We evaluated the system on both a European and national law corpus and analyzed its performance with respect to a baseline model and the Stanford NER tagger. Our results indicate that the concept recognition system is able to identify concepts in both directives and UK statutory instruments with a F1 score of 0.71 over the combined corpus. It can also be used to iteratively improve the dictionary-lookup based tagging from IATE. We also demonstrated that concept recognition systems are useful to align legal terminology at European and national level to assist legal practitioners and domain experts.

Acknowledgements

Research presented in this paper is conducted as a PhD research at the University of Turin and the University of Luxembourg within the Erasmus Mundus Joint International Doctoral (Ph.D.) programme in Law, Science and Technology. This work has been partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts". The authors would like to thank Ba Dat Nguyen from the Max-Planck Institute for his comments and suggestions.

References

[1] Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker. *16th International Conference on Artificial Intelligence and Law (ICAIL), 2017.*

[2] Xavier Carreras, Lluis Marquez, and Lluıs Padro. Named entity extraction using AdaBoost. *6th Conference on Natural language learning*, volume:20, pages:1-4, 2002.

[3] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. Named entity recognition and resolution in Legal text. *Semantic Processing of Legal Texts*, pages:27-43 Springer, 2010.

[4] Jenny Rose Finkel, Trond Grenager, and Christopher Manning Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages:363-370. 2005.

[5] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, pages:282-289, 2001.

[6] Rohan Nanda, Luigi Di Caro and Guido Boella. A Text Similarity Approach for Automated Transposition Detection of European Union Directives. *Proceedings of the 29th International Conference on Legal Knowledge and Information Systems (JURIX2016)*, pages:143-148, 2016.

[7] Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Dexter 2.0: An Open Source Tool for Semantically Enriching Data. *ISWC-PD'14 Proceedings of the 2014 International Conference on Posters and Demonstrations Track*, volume:1272, pages:417-420. 2014.

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

Subject Index

active machine learning 11
agent mentions 39
argument schemes 69
argumentation 107
argumentation support 69
automatic annotation 125
balancing 107
bar exam 89
bayesian probability theory 69
belief revision 49
case law 39, 95
case law analysis 177
case law databases 113
change detection 165
chinese legal documents 135
citation networks 59
CJEU 59
concept maps 21
concept recognition 193
conditional random fields 39, 177
consumer contract 145
contract compliance 33
deep learning 155
deontic rules 1
dimensions 27
European law 193
factors 27
foreeseability 171
formal semantics 33
full-text search 119
fuzzy time 33
goal-based reasoning 107
inadvertence 171
information extraction 135, 177
information retrieval 193
jurisprudences 119
justification 89
knowledge representation 135
LDA 21
legal analysis 39
legal analytics 183
legal case based reasoning 27
legal ontology 135
legal recommender system 21
legal references 113
legal responsibility 171
legal text analytics 155
legal text processing 101
linked data 1
linked open data 113
logic 89
machine learning 145, 165
markup language 101
methodology 101
moral responsibility 171
named entity recognition 39
natural language processing (NLP) 113, 119, 155
negligence 171
network analysis 59, 95
norm change 49
norm classification 11
OAB 89
ontology 1
Portuguese 119
question-answering 89
real-time contracts 33
reasoning about evidence 69
reference recognition 177
regulation 165
regulatory change management 165
risk 171
semantic annotation 101
semantic extraction 183
semantic modeling 125
semantic similarity search 125
semantic web 1
similarities 119
simplicity theory 171
statutory interpretation 107
temporal reasoning 49
text analytics 135
text mining 11
text similarity 59
topic clouds 21
topic model 135

unfair terms detection 145
value-based reasoning 107
values 27
vector space models 183
version 165
visual analytics 95

Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

Author Index

Agnoloni, T.	113
Al-Abdulkarim, L.	79
Androutsopoulos, I.	155
Araszkiewicz, M.	107
Arsuaga Lecuona, A.	113
Ashley, K.D.	39, 183
Asooja, K.	165
Atkinson, K.	27, 79
Bacci, L.	113
Bardelmeijer, S.	21
Bench-Capon, T.	27, 79
Boada García, A.	113
Boella, G.	193
Boer, A.	21
Bonczek, G.	11
Bougueng, R.	125
Bourguet, J.-R.	119
Bujor, O.	113
Cambronero, M.-E.	33
Casini, G.	v
Cervone, L.	113
Chalkidis, I.	155
Contissa, G.	145
Costamagna, F.	193
Cuconato, B.	89
Dasgupta, S.	135
Delfino, P.	89
Dessalles, J.-L.	171
Di Caro, L.	113, 193
Domhnaill, B. Ó	165
Falakmasir, M.H.	183
Foghlú, O. Ó	165
Gandon, F.	1
Glaser, I.	11
Gough, F.	101
Governatori, G.	1, 49
Gupta, A.	135
Haeusler, E.H.	89
Harašta, J.	177
Hong, H.	135
Lagioia, F.	145
Levy, F.	101
Liebman, B.L.	135
Lin, K.	135
Lippi, M.	145
Llana, L.	33
Lynch, M.	101
Marchin, G.	165
Martinez, D.C.	49
Matthes, F.	11
McGrath, S.	165
Micklitz, H.-W.	145
Muhr, J.	11
Nanda, R.	193
Nazarenko, A.	101
Nejadgholi, I.	125
Pace, G.J.	33
Palka, P.	145
Palmirani, M.	113
Panagis, Y.	59, 145
Peruginelli, G.	113
Prakken, H.	69
Rademaker, A.	89
Robaldo, L.	193
Roberts, M.E.	135
Rotolo, A.	49
Šadl, U.	59
Saillenfest, A.	171
Sartor, G.	145
Šavelka, J.	39, 177
Scepankova, E.	11
Sileno, G.	171
Siragusa, G.	113, 193
Stern, R.E.	135
Sun, H.	135
Tamargo, L.H.	49
Tarissan, F.	59
Theobald, M.	193
Torroni, P.	145
van den Oever, J.	113
van Dijck, G.	95
van Kuppevelt, D.	95
van Opijnen, M.	113
Villata, S.	1
Waltl, B.	11
Wang, A.Z.	135

Whittle, S. 79
Williams, R. 79
Winkels, R. 21
Witherspoon, S. 125
Wolfenden, C. 79
Wyner, A. v, 101
Zorzanelli Costa, M. 119
Zurek, T. 107