Domain-Driven News Representation Using Conditional Attribute-Value Pairs

Mihail Minev and Christoph Schommer

Interdisciplinary Lab for Intelligent and Adaptive Systems
Computer Science and Communications Research Unit
University of Luxembourg
{mihail.minev,christoph.schommer}@uni.lu

Abstract. Financial news carry information about economical figures and indicators. However, these texts are mostly unstructured and consequently hard to be processed in an automatic way. In this paper, we present a representation formalism that supports a linguistic composition for machine learning tasks. We show an innovative approach to structuring financial texts by extracting principal indicators. Considering announcements in the monetary policy domain, we distinguish between attributes and their values and argue that attributes are to be represented as an aggregated set of economic terms, keeping their values as corresponding conditional expressions. We close with a critical discussion and future perspectives.

Keywords: Feature Extraction, Text Representation, Financial News.

1 Introduction

Written text is the most common used format for the announcement of financial news. It usually includes several standard elements, in particular a title, a date, a location, an author, and a content, respectively. While a coherent text can normally be read and interpreted by humans (quite easily), an associative text analysis requires a complete and logical formal representation. Although many electronic publications feature structural metadata as well, such specifications do not cover the annotation of economic terms: moreover, these remain hidden in the text.

In this paper, we focus on a single class of documents, which is financial news related to *monetary policy* and being conducted by the *Federal Reserve*[1] (Fed). The corresponding news documents are provided by Thomson Reuters NewsScope. In particular, we examine the time period 2007–2012, which captures the development of the subprime mortgage crisis in the United States. We consider only official press releases, which are issued periodically by the Fed and which include the latest economic information as well as the arranged decisions by the *Federal Open Market Committee* (FOMC, the liable subdivision of the Fed).

[1] This is the central bank on the United States; its main goals are to monitor the price stability and to foster maximum output and employment [1].

N. Ferro (Ed.): PROMISE Winter School 2013, LNCS 8173, pp. 218–226, 2014.
© Springer-Verlag Berlin Heidelberg 2014

In general, financial news address a single company or an industrial branch, whereas Fed announcements have a significant impact on the entire [2] economy of a country. Typically, the FOMC announcements comply to a specific structure concerning the past and the current state of the economy as well as the promoted committee decisions. In addition, the documents comprise a high number of principal indicators, e.g., numbers for the labor market and the housing sector as well as the target range for the federal funds rate. Referring to the text composition a concrete example is available in Section 3.1.

The purpose of this work is to describe a novel approach for accumulating economic information in *monetary policy* news by incorporating linguistic aspects. First, we apply a shallow parser on the input and identify key attributes (features)[2]. We define an attribute as a *noun phrase* (NP), which is a phrase with a noun as a head word [3]. In this way, determined features are aggregated and ranked by frequency. Next, this feature list is likewise filtered and confirmed by financial domain experts. In a second step we obtain candidate *attribute values* (instances), which are either a *verb phrase* (VP), an *adverb phrase* (AdvP), or an *adjective phrase* (AdjP). For example, a valid combination is *'unemployment rate'* (attribute) and *'went down'* (value). Finally, the output is annotated for machine learning tasks, such as the analyzis of the correlations between the *monetary policy* decisions and particular stock market volatilities.

The paper is organized as follows: Section 3 provides a literature overview focusing on financial text representation methods. It also describes the model requirements in regard to the *monetary policy* domain along with the addressed linguistic aspects. In Section 4 we present an extraction and annotation model for the selected *attribute-value* pairs. We close with a critical discussion and prospective future works (Section 5).

2 Related Works

Studies concerning financial text representation distinguish between two main approaches: *unigrams* (or single terms) and *compositions* (or multi-word terms). Whereas *unigrams* are mainly used as individual, independent features, the second approach concerns features that are a composition of one or more words (following context-free grammars) and statistical measures. Since we emphasize on high information retention in the *monetary policy* domain, this is also the focus of the following literature survey.

2.1 Single-Word Terms

Representing documents as isolated words has been initially described by [4] as the *Vector Space Model*. The approach is also known as *Bag-of-Words* (BoW) and is preferred in many studies due to its ease of use. Among others, [5,6] apply this model to financial texts. A main requirement for the operation of a learning

[2] In this work the words *attribute* and *feature* are used as synonyms.

algorithm is that all feature constructs should be explicitly defined. Features are the linguistic counterpart of concepts in a particular domain. In the *Bag-of-Words* model, each of the features corresponds to a single term, which is assumed to be the meaningful unit in a sentence. Many comparative studies approved the high yields of *Bag-of-Words* for document classification [7,8]. However, BoW generates numerous (and noisy) features disregarding multi-word terms, which are typical for financial texts. Another main disadvantage of the single term representation is the information, which is discarded from the original text. The word order in sentences is not considered, but also the syntactic and semantic structures of word compounds are abolished. In spite of preserving the last, several alternative approaches are discussed in the next section.

2.2 Multi-word Terms

A second text representation method incorporates *multi-word terms* and – with them – domain affiliations and word relationships. This is why *multi-word terms* are a strong candidate technique for retaining as much semantics as possible. However, there is no evidence for a straightforward analogy between the length of a feature, which in this case is a compound, and the vocabulary in a particular domain [9]. In [10], a categorization scheme for *multi-word terms*, which is based on part of speech analysis, is suggested. According to the results, the most expressive compounds are the noun phrases, which also include the adjective-nouns and the phrasal compounds [9]. On the other hand, not all of these are terminology-relevant for the financial domain. Therefore, a thorough field expertise for the candidate assessment is still indispensable.

More lexical issues with multi-word term recognition are explained by [9]. The authors claim that even though a direct juxtaposition may indicate a terminology, it does not guarantee it. Composite terms can not be determined only by a set of formal syntactic rules. The English language structures and parsing ambiguities inhibit a clear distinction between *multi-word terms* and general language compounds. Variations like hyphenations and abbreviations hinder likewise the process and hence the comparability of the results. In contrast, *n-grams* are a language independent method for text representation. Word bi- and tri-grams are a popular topic of recent scientific studies, though some specific research focused also on character and byte n-grams [11]. Word n-grams are defined as adjoined strings in texts. To reduce the number of candidates, which may be huge for a large document set, a number of statistical measures may be applied: Likelihood ratio, Chi-square, and Pointwise Mutual Information. Besides the high dimensionality, n-grams lack the complete representation power, which noun phrases have [12]. Obviously due to the non-consideration of lexical structures. The information they capture is insufficient and can be used only partially (in combination with some other method) for an extended text abstraction.

In a comparative study [13] examines three widespread financial news representation techniques including Bag-of-Words, Noun Phrases and Named Entities. The latter is an extension of Noun Phrases, which assigns a particular category to a subset of its terms like date, location, money, organization, percentage,

person, and time. The evaluation upon a stock price prediction task were ambiguous as none of the methods dominated the field. Nevertheless Noun Phrases achieved the best results in two out of tree prediction metrics. Further attempts for an independent term selection and their ranking include the application of statistical methods as Mutual Information [6] and Chi-Square [14]. Nevertheless, none of the previous studies concerns the discrete relationships between the extracted features and their instances. In the next section we introduce the model requirements for such an approach.

3 Model Requirements

The model requirements are indicated by two aspects. The first aspect involves the understanding of the *monetary policy* domain and the corresponding press releases. The second aspect concerns the study of the content and the structure of the documents as well as the formal identification of the candidate attribute-value pairs.

3.1 Domain Understanding

Monetary policy news are packed with indicators, e.g., the recent developments on the labor market, the average interest rates, and the latest economic barometers. Accordingly, our goal is to transform such stories into a structured format using feature-value pairs, which enclose one or more coherent words. Logically, the first step is to examine the official press releases (issued by the Federal Reserve) in a concrete time period. To demonstrate, we have chosen a time interval between 2007 until 2013, because the year of 2007 has been recognized by many officials [15] as the begin of the *subprime mortgage crisis* in the United States. In the six year time interval, 55 FOMC statements (21520 tokens) for the tracking the eminent *monetary policy* are considered.

The *Federal Open Market Committee*, which is responsible for setting the *monetary policy*, meets regularly eight times per year. Subsequently, the committee releases an official statement; afterwards the chairman of the Federal Reserve, Ben Bernanke, stages a question and answer session. The released records comprise the latest principal indicators, but also the short-term economy expectations as well as proposed measures for interventions. Evidently, the conclusions disclosed by the *Federal Open Market Committee* have a significant influence on the entire US industry [2] and as a result the information is greatly anticipated by politicians and investors.

In this work we describe a method to extract the facts from the Fed announcements and to enable dedicated economic surveys, for instance—an analysis of the correlations between the federal funds rate and the unemployment rate; or between the asset purchase programs and the stock markets. In order to do so, we first need to quantify the information before we can apply learning algorithms to identify associated patterns. We begin with examining the structure and the content of a random FOMC document (June 20, 2012). Here is a snippet from the first paragraph:

> "Business fixed investment has continued to advance. Household spending appears to be rising at a somewhat slower pace than earlier in the year. Despite some signs of improvement, the housing sector remains depressed. Inflation has declined, mainly reflecting lower prices of crude oil and gasoline, and longer-term inflation expectations have remained stable."

Clearly, the fragment provides information about the current economy state and includes at least one principal indicator per sentence (marked below in bold). The next part of the message is devoted to the committee expectations such as the *labor market conditions*, *long-term inflation*, and *economic growth*. The third and fourth paragraphs give information about the *Federal Funds Target Rate*, the *Maturity Extension Program* (alias *Operation Twist*) and the *Asset Purchase Program* (alias *Quantitative Easing*). The attendees' names and their voting *pro/against* the proposed measures come at the end. Empirically, all *Federal Open Market Committee* announcement within the examined period share similar structure:

> *"**Business fixed investment** has continued to advance. **Household spending** appears to be rising at a somewhat slower pace than earlier in the year. Despite some signs of improvement, the **housing sector** remains depressed. **Inflation** has declined, mainly reflecting lower prices of **crude oil and gasoline**, and **longer-term inflation expectations** have remained stable."*

As mentioned previously, our objective is to create a method for the identification and extraction of attributes, which are equal to economic indicators. In line with our analysis, the attributes are noun phrases, with a high frequency distribution over the text collection. Furthermore we look for their conditional values in the text, which are labeled consistent with syntactic rules..

3.2 Linguistic Processing

Each document in our collection is annotated corresponding to the process shown in Fig. 1 [16]. In the first part of the workflow, we carry out the sentence segmentation, the tokenization and the part-of-speech (POS) tagging. The output is a set of tuples, where each word is assigned to a lexical class, e.g., (investment, NN). The tag NN hereby refers to a *noun, singular or mass*. This is a standard preprocessing step before the phrasal category detection (also *chunking*). Second, a *maximum entropy*-based chunker evaluates the input pairs and assigns labels to each syntactic word group. Currently, four types of phrases are detected: *noun phrase*, *verb phrase*, *adverb phrase*, and *adjective phrase*. For example, the construct [business, NN fixed, VBN investment, NN] is a noun phrase. In this case it is a multi-word attribute with a verb in past participle surrounded by two singular nouns. All syntactic rule definitions for the phrase annotation are summarized by [17].

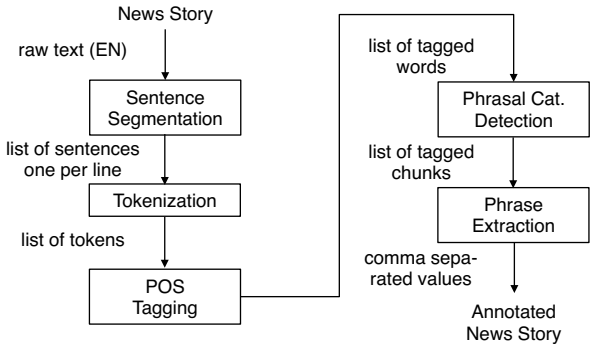

Fig. 1. This workflow describes how the document annotation process works: starting from a news story, the text is firstly linguistically preprocessed (Sentence Segmentation, Tokenization, Part-Of-Speech tagging). Next, the phrasal category detection is completed using a trained model. The annotated news story is received by filtering the four tagged chunks (NP, VP, AdvP, AdjP).

In the extraction phrase, we separate the chunks (per sentence) and file them as comma separated values. The output accumulates all identified phrases in a document. Based on our experiments with the training data, four words per phrase are not exceeded. In order to determine the attributes (economic indicators) from the data, we collect all noun phrases (in total 705). According to [18,19], the noun phrase is the most expressive construct in a sentence, ergo suitable as a candidate for a domain vocabulary. However, our list aggregates also personal and location names as well as roughly 15% incorrectly identified noun phrases. In order to trim the candidate attributes we rank the NPs using the *C-Value* [20,21] algorithm, which incorporates assorted frequency measures. Despite some improvements[3] in the feature distribution, the false positive values remain high. External domain knowledge is, therefore, explicitly needed, which evokes us to ask financial experts to select those noun phrases, which represent economic indicators. Based on their votes, we aggregate all matching attributes to a domain specific vocabulary T (here, 153 unique records are listed).

With respect to the linguistic annotation, we utilized the *OpenNLP* machine learning toolkit [22]. Due to the data similarity we measured the *OpenNLP* chunking performance on the *CoNLL-2000* [23] test set with 47377 tokens. *CoNLL-2000* contains syntactically annotated sentences from *'The Wall Street Journal'*. The results for precision and recall were respectively 0.93% and 0.92%.

4 Attribute-Value Representation

For the text representation, we extend a formalism initially proposed by [24] for improving the retrieval performance in search queries. The author describes

[3] Setting the *C-value* rate to one and above, which is a typical threshold, reduces the number of noun phrases to 509.

a method for parsing search strings by adopting their semantic and syntactic features. The study assumes, that the queries are not expressed as full sentences, but built up of distinct nouns and/or noun phrases. Which is a divergence to the grammatically correct texts and the four lexical phrases we encounter.

Following the model requirements as presented in Section 3, we acquire a set of annotated documents. Consequently, we define an attribute a as

$$\{a \mid a \in T \ \wedge \ P(a)\}\,. \tag{1}$$

where T stands for the vocabulary we use. $P(a)$ is true if and only if a set of terms exists, which satisfy a condition a [16]. In this case, a must be a noun phrase, as requiered for the domain specific vocabulary. For the attribute values av we apply the definition

$$\{av \mid av \in P(av)\}\,. \tag{2}$$

Here, the property $P(av)$ is true if and only if av is either a verb phrase, or an adverb phrase, or an adjective phrase. For each *attribute* per sentence the matching *attribute values* are retrieved. Accordingly, we outline a representation schema, which incorporates the three integrals:

1. *attribute* $\in T$;
2. *attribute value* $\in (VP \vee AdvP \vee AdjP)$;
3. a class C, which describes the time-variant type of the *attribute value*.

We can determine the time frame C based on the token's POS tags, which are available for the *verb phrases*. Each *attribute* has one or more expressions of an *attribute value*. An *attribute value* exists only in a combination with an *attribute* and is assigned to zero or one class C. In a composition, an *attribute* and an *attribute value* establish an unique pair for each sentence. For example, the sentence

> *"However, growth in employment has slowed in recent months, and the unemployment rate remains elevated."*

is annotated with the values

[*attribute*: employment]
[$attribute\_value_{past\_state}$: has_slowed]
[*attribute*: unemployment_rate]
[$attribute\_value_{present\_state}$: remains]
[*attribute_value*: elevated].

For longer sentences the identification of the attribute instance(s) can be ambiguous. To avoid redundant value allocations we apply syntax-based rules, which use lexical delimiters like a comma or a dash and support partitioning. In the latter example we split the sentence in three parts (delimiter is a comma) and determine the attribute-value pairs in each case.

5 Conclusions

A strong limitation of our approach is its domain dependency. Although practicable for the financial text representation, it correlates also with contributions in other areas like medicine (see [25]). In the same context, its application on more generic news is still challenging: this is because of the natural ambiguity and the linguistic complexity of universal texts. Besides that, for a more precise *attribute-value* identification we also plan to conduct experiments using dependency trees.

A quantitative model evaluation is targeted in a future work. Typical text classification measures like *Precision/Recall* are barely applicable due to the specific attribute-value format. Clustering and similarity techniques sound more promising, though further definitions of the comparability parameters are necessary. At present, our alternative idea is to add time series data and to juxtapose the prediction results of stock market trends with analog studies.

In this work, we have proposed a fresh approach for representation of *monetary policy* news in the context of machine learning applications. In order to quantify a Federal Reserve document, we have considered the lexical structure of the texts as well as the semantic relationships between the terms. In this context, we acquire a set of four phrase types, which enable multi-word term identification in financial texts. Correspondingly, we have designed an annotation model to capture the domain-specific information as conditional attribute-value pairs. One future application of this work is to facilitate economic surveys. For example, we may track the *monetary policy* implementation over a dedicated time period and/or measure the correlations between principal indicators, e.g., the policy instruments, the various interest rates and the economy state.

References

1. Federal Reserve Bank of New York, `http://www.newyorkfed.org/aboutthefed/fedpoint/fed48.html`
2. Bernanke, B.S., Kuttner, K.N.: What explains the stock market's reaction to federal reserve policy? Working Paper 10402. National Bureau of Economic Research (April 2004)
3. Radford, A.: English Syntax: An Introduction. Cambridge University Press (May 2004)
4. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (1975)
5. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J.: Language models for financial news recommendation, pp. 389–396 (2000)
6. Gidófalvi, G., Elkan, C.: Using news articles to predict stock price movements. Technical report, Department of Computer Science and Engineering, University of California (2003)
7. Fengxi, S., Liu, S., Yang, J.: A comparative study on text representation schemes in text categorization. Pattern Anal. Appl. 8(1), 199–209 (2005)
8. Scott, S., Matwin, S.: Feature engineering for text classification. In: Proceedings of ICML 1999, 16th International Conference on Machine Learning, pp. 379–388. Morgan Kaufmann Publishers (1999)

9. Frantzi, T.K., Ananiadou, S.: Automatic term recognition using contextual cues. In: Proceedings of 3rd DELOS Workshop (1997)
10. Sager, J.C., Dungworth, D., McDonald, P.F.: English special languages: principles and practice in science and technology. Brandstetter (1980)
11. Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J., Spiteri, R.: Document representation and dimension reduction for text clustering. In: ICDE Workshops, pp. 770–779 (2007)
12. Radford, A.: Syntactic Theory and the Structure of English: A Minimalist Approach (Cambridge Textbooks in Linguistics). Cambridge University Press (August 1997)
13. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans. Inf. Syst. 27(2) (2009)
14. Hagenau, M., Liebmann, M., Hedwig, M., Neumann, D.: Automated news reading: Stock price prediction based on financial news using context-specific features. In: Hawaii International Conference on System Sciences, pp. 1040–1049 (2012)
15. Federal Reserve Bank of St. Louis, `http://timeline.stlouisfed.org/pdf/CrisisTimeline.pdf`
16. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python, 1st edn. O'Reilly Media (July 2009)
17. Marcus, M., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The penn treebank. Computational Linguistics 19(2), 313–330 (1993)
18. Schwarzschild, R.: The role of dimensions in the syntax of noun phrases. Syntax 9(1), 67–110 (2006)
19. Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M.: Terminology extraction: An analysis of linguistic and statistical approaches. In: Sirmakessis, S. (ed.) *Knowledge Mining*. STUDFUZZ, vol. 185, pp. 255–279. Springer, Heidelberg (2005)
20. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the c-value/nc-value method. Int. J. on Digital Libraries 3(2), 115–130 (2000)
21. Java Automatic Term Extraction toolkit, `http://code.google.com/p/jatetoolkit/wiki/JATEIntro`
22. Apache OpenNLP library, `http://opennlp.apache.org/`
23. CoNLL-2000, `http://www.cnts.ua.ac.be/conll2000/chunking/`
24. Li, X.: Understanding the semantic structure of noun phrase queries. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 1337–1345. Association for Computational Linguistics, Stroudsburg (2010)
25. Milios, E., Zhang, Y., He, B., Dong, L.: Automatic Term Extraction and Document Similarity in Special Text Corpora, pp. 275–284 (August 2003)