nel, of co-speech gesture, eye gaze, facial expressions, posture and possibly other kinetic/visual channels, combinations of image and text (Dancygier and Vandelanotte 2017:567). In order to enhance functions of AI, this study argues that AI-interfaced humanities start in a full scale modeling the construing process of the multimodal data. In the vein, it introduces qualitative analyses of multimodal data to the field and initiates to discuss the possibility of designing processing of multimodal data within the AI-interfaced humanities framework.

References

Lakoff, George and Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to western thought.* New York: Basic Books.

Dancygier, Barbara and Lieven Vandelanotte. 2017. Internet memes as multimodal constructions. *Cognitive Linguistics, 28*(3):565-598.

Mind and Language:

AI in an Example of Similar Patterns of Luxembourgish Language

Joshgun Sirajzade, Christoph Schommer
University Luxembourg

Is the language the key to the mind? Or does the mind determine our ability to speak? Being one of the oldest questions of humanities (especially in philosophy and linguistics) and science (later social and computer science) it still remains unsolved. Yet, in the last decades there has been huge successes in the field of NLP and Computational Linguistics on the one hand, on the other hand the current research shows that not only the human mind can shape our language, but also the language can program our mind [1]. Advances in Speech Recognition, Text Processing and Language Generation brought the Human Machine Communication to a new level. In this talk we suggest a new language theory called text flow in an example of Luxembourgish language and compare it against existing language models (e.g. probabilistic language model). We also take into account the role of concrete steps of NLP in this model (like Tokenization, Normalization, POS-Tagging, Stemming, Lemmatization or other methods for morphological and syntactical segmentation). This new theory considers language laws as described in quantitative linguistics [2], such as Zipf-Mandelbrots law or token-type-ratio etc. Furthermore, we discuss our model in the context of existing techniques of machine learning and knowledge representation because the applicability of the suggested theory is the main focus of our research. Nevertheless, this model can help us to achieve a deeper understanding of a particular language it is applied to. The model and concrete examples from its application in our talk are driven from a corpus of Luxembourgish language. It consists of ca. 130 mio. tokens being mostly news from the web presence of Radio Television of Luxembourg (RTL) with ca. 35 mio. tokens, Radio news transcriptions (RTL) with ca. 20 mio. tokens, over 70 mio. tokens user commentaries (also on the RTL web presence), some legal texts and parliament speeches from Luxembourgish Parliament with ca. 10 mio. tokens and interview transcriptions representing a spoken language. The new theory considers syntagmatic and textual language structures as "flowing signals" [3, 4]. It says, that a language signal, that is uttered, have different formal properties like the length

of the sequence (words, sentences or even larger text pieces) and the order of the items in these. The mathematics behind the theory takes into account the repetition of similar patterns in the sequences. Here, the signals of different size and order can be semantically and pragmatically related or even equal, meaning there are more than one way to say things, whereas the two or more sequences, which are formally the same, can only mean the same thing. Morphologically, this model does not distinguish function words from affixes. They are all repeating patterns in the language system and identifying and generalizing them allows to extract the grammar of one particular language.

Key words:
(ACM Classification): I.2.7 Language Models, Text Analysis; I.2.0 Cognitive simulation; I.5 Pattern Recognition; J.5.5 Linguistics;

References

Lupyan, G. & Bergen, B. (2015). How language programs the mind. *Topics in Cognitive Science.* New Frontiers in Language Evolution and Development. 10.1111/tops.12155.

Hřebíček, Luděk (2005). Text Laws. In *Quantitative Linguistics: An International Handbook,* eds. Köhler, Reinhard., Gabriel Altmann a Rajmund G. Piotrowski. Berlin, New York: de Gruyter, 348-361.

Joshgun Sirajzade (2016). *Compiling Tools and Resources for Studying of Luxemburgish Language and beyond.* DHBenelux Conference, Luxembourg.

Joshgun Sirajzade (2018). Korpusbasierte Untersuchung der Wortbildungsaffixe im Luxemburgischen. Technische Herausforderungen und linguistische Analyse am Beispiel der Produktivität. in Zeitschrift für Wortbildung = *Journal of Word Formation, 2018*(1).

생각과 언어:

룩셈부르크어의 유사한 패턴의 예에서의 AI(인공지능)

Joshgun Sirajzade, Christoph Schommer
룩셈부르크 대학 CSC 학과 ILIAS 연구실

언어는 생각의 열쇠인가? 아니면 생각이 우리의 말하기 능력을 결정하는가? 이는 인문학(특히 철학 및 언어학)과 과학(후기 사회과학 및 컴퓨터 과학)의 가장 오래된 질문 중 하나로서 여전히 풀리지 않은 채 남아 있다. 그러나, 지난 수십 년 동안 자연어 처리(NLP) 분야 및 컴퓨터 언어학 분야에서 큰 성공을 거둔 반면, 현재의 연구 수준에 비추어볼 때, 인간의 생각이 우리의 언어를 형성할 수 있을 뿐만 아니라, 언어가 우리의 생각을 *프로그래밍*할 수 있다[1]. 음성 인식, 텍스트 처리 및 언어 생성 기술이 발전함에 따라 휴먼 머신 커뮤니케이션의 수준이 새로운 수준으로 접어들었다. 이 강연에서는 룩셈부르크어의 예를 들어 *텍스트 흐름(text flow)*이라는 새로운 언어 이론을 제안하고, 확률론적 언어 모델 등과 같은 기존의 언어 모델과 비교하도록 한다. 또한 이 모델에서 자연어 언어(NLP)의 구체적인 단계의 역할을 고려한다(토큰화, 정규화, POS-태깅, 스테밍(Stemming, 어간추출), Lemmatization(표제어 찾기) 등) 기타 형태학적 구문론적 분류 방법). 이 새로운 이론은 Zipf-Mandelbrots law 나 token-type-ratio 등, 정량적 언어학[2]에 기술된 언어 법칙을 고려한다. 또한 제안된 이론이 적용성이 우리 연구의 주요 관심사이기 때문에, 이 모델을 머신 러닝 및 지식 표현의 기존 기술의 맥락에서 논의하고자 한다. 그럼에도 불구하고, 이 모델은 이 모델이 적용되는 특정 언어에 대해 보다 심도있게 이해할 수 있도록 도움을 줄 것으로 기대한다. 본 강연에서 적용된 모델과 구체적인 예들은 룩셈부르크어의 언어자료로부터 비롯되었다. 그것은 주로 룩셈부르크 라디오 텔레비전(ca. 35 mio. token), 캘리포니아의 라디오 뉴스 기록 (RTL)(ca. 20 mio. tokens), 사용자 주석 (70 mio. tokens 이상)(또한 RTL 웹 현실화에서), 일부 법률적 텍스트와 룩셈부르크 의회의 의회 연설 (ca. 10 mio. tokens) 및 구어체를 사용하고 표현하는 인터뷰 발췌문 등의 웹 현실화 뉴스에서 온 ca. 130 mio. tokens으로 구성되어 있다. 새로운 이론에서는 통합적 언어 구조와 텍스트적 언어 구조를 "흐름의 신호(flowing signals)"로 간주하고 있다[3, 4]. 말로 표현되는 언어 신호는 시퀀스의 길이(단어, 문장 또는 심지어 더 큰 텍스트 조각)와 이들 안에 있는 항목의 순서 등과 같은 형식적 속성이 다르다. 그 이론의 근간이 되는 수학에서는 서열에서 유사한 패턴의 반복을 고려한다. 여기서, 다양한 크기와 순서의 신호는 의미론적으로 관련이 있을 수도 있고, 실용적인 면에서는 심지어 동일할 수 있다. 즉, 형식적으로 동일한 둘 이상의 시퀀스가 동일한 것을 의미할 수 있는 반면, 여러 가지 방법으로 표현할 수 있다는 것을 의미한

다. 형태학상 이 모델은 기능어와 접미사를 구분하지 않는다. 그것들은 모두 언어 체계에서 반복되는 패턴이며, 그것들을 식별하고 일반화함으로써 하나의 특정한 언어의 문법을 추출할 수 있다.

키워드:
(미국컴퓨터학회(ACM) 분류): I. 2.7 언어 모델, 텍스트 분석, I.2.0 인지적 시뮬레이션, I.5 패턴 인식, J. 5.5 언어학,

참고문헌

Lupyan, G. & Bergen, B. (2015). 언어가 생각을 프로그래밍하는 방법 *인지 과학의 주제.* 언어의 진화와 발전에 있어서의 새로운 개척자 10.1111/tops.12155.

Hřebíček, Luděk (2005). 텍스트 법칙. *양적 언어학 : 국제 핸드북,* eds. Köhler, Reinhard., Gabriel Altmann a Rajmund G. Piotrowski. Berlin, New York: de Gruyter, 348-361.

Joshgun Sirajzade (2016). *룩셈부르크어 및 그 외의 언어를 연구하기 위한 컴파일 도구 및 리소스.* DHBenelux 학술회의, 룩셈부르크

Joshgun Sirajzade (2018). Korpusbasierte Untersuchung der Wortbildungsaffixe im Luxemburgischen. Technische Herausforderungen und linguistische Analyse am Beispiel der Produktivität. in Zeitschrift für Wortbildung = *Journal of Word Formation, 2018*(1).

Session 9