Set-Based Modeling of Ambiguous Word Embeddings using Deep Learning

Background

In natural language processing, one of the most effective ways to represent words is through word embeddings, where each word is represented as a vector. Text corpora and neural networks are used to train these vectors using variations of the word2vec algorithm [6], which trains each word based on its surrounding words. After training, these word embeddings can be utilized for further processing, with large language models being a prominent example [7].

However, a significant issue arises with homonyms: words that have multiple meanings [8]. In traditional word embeddings, these ambiguities are represented in a single vector, which results in an averaged meaning of a given word. However, nuances or even contradicting meanings of a word are lost using this approach.

One approach to addressing this issue is to detect and replace homonyms [5]. However, these methods typically detect homonyms using a rule-based or statistical-based approach, which requires an answer set and expert knowledge of the analyzed text.

Description

The goal of this thesis is to develop a model that automatically captures multiple meanings of words using set-based computing. Thus, we model the word embeddings not as single vectors but as continuous sets, e.g., a zonotope [3]. As these sets can easily be parametrized [4], a meaning of the word can later be extracted from the set depending on the current context. This method is expected to represent the different meanings of a single word, potentially providing a more nuanced and accurate depiction of its semantic range.

To achieve this, the ambiguous word embeddings will be learned by adapting word embedding algorithms to output a zonotope, which consists of a center and several generators. The generators will allow the meaning of a word to be represented by a set rather than a single vector, effectively capturing the different meanings of homonyms even in high-dimensional embedding spaces.

Tasks

- Literature research on ambiguous words in natural language processing
- Familiarize with the toolbox CORA [1]
- Dataset selection and preprocessing
- Implementation of set-based word embeddings
- Evaluation on the selected data set
- Optional: Explore more complex set representations, e.g., polynomial zonotopes.

References

[1] Matthias Althoff. An introduction to cora 2015. In *ARCH@ CPSWeek*, pages 120–151, 2015.

[2] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 2019.

[3] Antoine Girard. Reachability of uncertain linear systems using zonotopes. In *International workshop on hybrid systems: Computation and control*, pages 291–305. Springer, 2005.

[4] Niklas Kochdumper, Bastian Schürmann, and Matthias Althoff. Utilizing dependencies to obtain subsets of reachable sets. In *Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control*, pages 1–10, 2020.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Tobias Ladner, M.Sc.

Research project:
FAI

Type:
BT

Research area:
Natural language processing, set-based computing

Programming language:
Python

Required skills:
Knowledge in formal methods and machine learning, good mathematical background

Language:
English

Date of submission:
24. Mai 2024

For more information please contact us:

Phone: +49 (89) 289 - 18140

E-Mail: tobias.ladner@tum.de

Website: ce.cit.tum.de/cps/

[5] Younghoon Lee. Systematic homonym detection and replacement based on contextual word embedding. In *Neural Processing Letters*, pages 17–36, 2020.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[8] Apurwa Yadav, Aarshil Patel, and Manan Shah. A comprehensive review on resolving ambiguities in natural language processing. *AI Open*, 2:85–92, 2021.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial
Intelligence and Real-time
Systems