University of Ljubljana
Faculty of Computer and
Information Science

FRI Natural language processing course 2021/2022
Project report

Cross-lingual sense disambiguation

Matej Miočić, Marko Ivanovski, Matej Kalc

Abstract

Word embeddings are unable to model the property of words to have potentially multiple meanings. Consequently, they cannot capture the particular meaning in which such a word has been used. To address this problem, sense or contextualized embeddings have been proposed for the English language. We will try to tackle this problem for the Slovene language.

Keywords
NLP, sense disambiguation, BERT, web scraper, tf-idf, corpora, classification

Advisors: Slavko Žitnik

Introduction

Like other languages, the Slovenian language has loads of polysemous words which appear in sentences having different meanings such as ”gol” and ”klop”. Such ambiguity can be a problem for many different NLP tasks resulting in a need for a model capable of understanding in which context the homonym appeared. This project is a simplified version where the main task is to determine whether the word is used in the same sense in both sentences. In this report, we will give initial ideas for the data acquisition and corpus preparation process. Finally, we will finish off by proposing approaches for the classification part of this task.

Related work

A similar problem has been solved for the English language. Wang et al. [1] created a large-scale Word in Context dataset, called WiC using sentences from different corpora. Each instance contained a target word, either a verb or a noun, for which two contexts were provided. Their task was the same as ours, to identify if the occurrences of a word in the contexts corresponded to the same meaning or not. The whole dataset was annotated by experts and those labels were used as ground truth.

After creating the dataset they experimented with recent word embedding techniques such as Context2Vec, ELMo and BERT. For classification, they used simple binary classifiers such as dense network and a threshold based on the cosine distance of the two input vectors. Their best model was BERT providing around 15.5% absolute improvement over a random baseline. Since the dataset is relatively small, the threshold-based strategy proved to be more efficient.

Methods

This task can be divided into two parts:

1. Corpus preparation
2. Classification and analyses

Corpus preparation

For the first part of our task, we first need to find as many polysemous words in the Slovenian language as possible and make a list of them. We have come up with two ideas on how we can extract such words:

- Our first idea is a semi-automatic approach which involves web scraping Slovenian’s dictionary web page ”fran.si”. For each Slovenian word, the website contains entries with the meanings of that word. The idea is to write our web scraper which would be initialized with the 10000 most frequent Slovenian words. Then, for each word it would check if it has more than one meaning.
- Our second idea is manually creating a list of homonyms using [2].

Once we obtain a list of these homonyms we can proceed to extract sentences that contain these polysemous words. For this, different Slovenian corpora can be used such as GigaFida [3] or many others from ”clarin.si”. After extracting and grouping all sentences in which our keywords appear the corpus is ready to be made.

Classification and analyses

After obtaining the corpus the analyses can be started. We think that this task can be solved using supervised and unsupervised methods. In order to work with any ML model sentences would need to be represented using some kind of embedding. Simple representations that can be tried are a bag of words, latent semantic analysis, or tf-idf weighting based on the documents in the corpus. Next, we can potentially try and cluster the sentences containing the same word into two clusters. If the clusters turn out to be good in dividing different context sentences we have potentially solved the problem. For classification, we can save the centroids of both clusters for every word. Then, when a new pair of sentences come on the input we can simply calculate the distance to those centroids. If both sentences turn out to be equally far from both centroids the word is probably used in the same context. Otherwise, one sentence would be closer to one centroid and the other sentence to the other one. Some threshold would need to be defined here, which would strictly differentiate both scenarios. The second supervised approach needs manual data annotation, which is highly time-consuming, or we can annotate semi-automatically using the clustering approach. After annotation we could try and train different models such as: word2vec [4], ELMo [5], BERT [6] and BERT++. For classification, we would use cosine similarity on the trained dense embeddings similarly to [1].

Conclusion

NLP disambiguation tasks aren't trivial, especially when we have no corpora at our disposal. However, with our semi-automatic process for creating the corpora, we believe we can try many different approaches for solving it. We have proposed different ideas for approaching such a task, but additional testing and experimenting would be required to come up with the most optimal approach, which maybe has not yet been mentioned. All that would follow, after we have the dataset to perform the analysis on.

References

[1] Mohammad Taher Pilehvar and Jose Camacho-Collados. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*, 2018.

[2] Júlia Bálint Čeh. *Slovar slovenskih homonimov: na podlagi gesel Slovarja slovenskega knjižnega jezika*. Znanstveni inštitut Filozofske fakultete, 1997.

[3] Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. Gigafida 2.0: The reference corpus of written standard slovene. In *Proceedings of the 12th language resources and evaluation conference*, pages 3340–3345, 2020.

[4] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[5] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.