University *of Ljubljana*
Faculty *of Computer and Information Science*

Natural language processing course Project proposal report (Task 2)

Lara Kuhelj, Gregor Novak, Jan Pelicon, Tomaž Štrus

Abstract

We will construct a tool that extracts hypernym-hyponym pairs from a domain-specific corpus. Furthermore we will construct another a tool for extraction of new instances of domain-specific semantic relations from the provided Karst corpus, using some of existing state of the art methods and algorithms such as KGPool, predicted method usage, BERT, Word2Vec.

Keywords

Karstology, semantic relation, hyponym-hypernym, extraction.

Advisors: Slavko Žitnik

Introduction

Even nowadays, problems that arise from natural language processing don't have a simple solution because of natural language complexity. Statistical methods that have been used previously to tackle such problems are being replaced by machine learning methods and deep neural networks that are now considered state of the art. Our goal is to use of this methods and train them on a few selected domain-specific corpuses in order to extract hypernym-hyponym pairs. Also, based on the existing instances of domain-specific semantic relations, we should be able to discover new instances semantic relations from provided Karst korpus, with a focus on a few chosen relations.

The Karst corpus

The korpus contains relevant works in kastrology. This includes books, reference works, PhD theses and scientific articles. It is provided as in TXT and in the ClassLa tagged format. From the paper we can see the size of the corpus in the table 1. As we do not have the source code for the transformation, we will use the ClassLa pipeline from GitHub [1] [2].

Table 1. Size of corpus by language

	English	Slovene	Croatian
Tokens	2,386,075	1,208,240	1,229,368
Words	1,968,509	987,801	969,735
Sentences	87,713	51,990	53,017
Documents	54	60	43

Related work and projected work plan

When it comes to relation extraction from text, various problem domains exist, with several methods and algorithms used for the purpose of solving them. Below we shortly describe a few different methods and how we plan to use them to solve the given tasks.

KGPool

KGPool is one of the newer methods, presented in 2021 [3]. It uses knowledge graphs for single sentence mapping with purpose of different relation extraction. It also incorporates neural methods. It is reported as one of the best performing methods on WikiData dataset and New York Times Corpus, which are often used for performance measuring when it comes to relation prediction [4].

Predicted method usage

We can use method for generating knowledge graphs on given domain and try to extract different context relation in a way, similar to the one shown in the Figure 1.

Word2Vec

Word2Vec [5] is one of the widely spread methods in natural language processings. It is a group if two-layer neural networks, trained for reconstructing linguistic context of words. As name suggests, method represents each word with unique

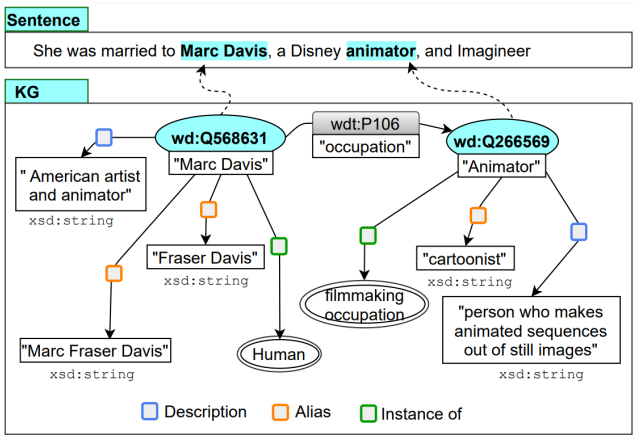

Figure 1. Example of knowledge graph [3]. We can see the extracted relation between *Marc Davis* and *Animator* (*occupation*).

vector, producing a vector space of several hundred dimensions. Words with common contexts in the corpus are located close to one another.

Predicted method usage

Method could be used for extracting hypernym-hyponym pairs if proper constraints are used for determining such pairs in generated vector space.

BERT

Bidirectional Encoder Representations from Transformers (BERT) is a NLP machine learning method based on transformers, proposed by Google [6]. Method is designed to pre-train deep bidirectional representations from unlabeled text. When pre-training is executed, method can then be fine-tuned and adapted for various specific tasks. Two BERT-based models were used in SemEval 2010 Task 8, both being one of the best performers. The task was as follows [4]: Given a sentence and two tagged nominals, to predict the relation between those nominals and the direction of the relation.

Example:

- "There were apples, **pears** and oranges in the **bowl**."
- Relation: (content-container, pears, bowl)

Predicted method usage

If similar approach is used in our task, we could try to find all of the nominals, where relation of the nominals is hypernym-hyponym, and determining which of the nominals is which with finding the direction of the mentioned relation as well.

Conclusion

We plan to solve given tasks with mentioned methods in a way, described in previous section. We will also compare different methods and try to determine which model performs the best both on the given corpus and other related problem domains. If possible, we will also try to combine results of different methods to achieve better accuracy.

References

[1] ClassLa code. Available at: https://github.com/clarinsi/classla. [Accessed: 18. 3. 2022].

[2] Nikola Ljubešić and Kaja Dobrovoljc. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy, August 2019. Association for Computational Linguistics.

[3] Abhishek Nadgeri, Anson Bastos, Kuldeep Singh, Isaiah Onando Mulang', Johannes Hoffart, Saeedeh Shekarpour, and Vijay Saraswat. KGPool: Dynamic knowledge graph context selection for relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 535–548, Online, August 2021. Association for Computational Linguistics.

[4] Relationship Extraction | NLP-progress. Available at: http://nlpprogress.com/english/relationship_extraction.html. [Accessed: 18. 3. 2022].

[5] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.