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

Natural Language Processing - Final Report (Task 2)

Lara Kuhelj, Gregor Novak, Jan Pelicon, Tomaž Štrus

Abstract

The task of automatic extraction of semantic relationships and definitions has become easier in recent years with the popularity and success of the deep learning models, such as BERT which stands for bidirectional encoder representations from transformers. In our case, we used two different BERT architectures and trained few models for hypernym-hyponym extraction and extraction of other semantic relations. Our models were finetuned on the TermFrame dataset of Karstology provided by the linguists that worked with us in cooperation. We also used the relations from SemRelData dataset for additional comparison and evaluation.
The models successfully learned and were able predict the selected relations with good accuracy and overall score, even when there were plenty of different clasification classes. We also tried two different ways of automatic entity recognition that are then used for relationship extraction - spaCy and entity recognition with custom rules that were based on empirical observations of distribution of words and phrases. Both approaches fell short of our expectation. While the spaCy model were in some cases able to approximately locate wanted entities it often included other words that messed with the relationship extraction.

Keywords

Karstology, domain-specific, semantic relation extraction, hyponym-hypernym extraction.

Advisors: Slavko Žitnik

Introduction

Even nowadays, problems that originate from the field of natural language processing, usually don't have a simple solution, because of natural language's complexity and its complex word and sentence formation. In the past, there were mostly statistical methods being used to tackle such problems, but nowadays, they are being replaced by machine learning methods, deep neural networks and transformers that are now considered state of the art.

Our goal was to use one or more of these methods and models, train them on a few selected domain-specific corpora and extract hypernym-hyponym pairs. Moreover, based on the existing instances of domain-specific semantic relations, we should be able to discover new instances of semantic relations from provided Karst corpus and other explored datasets and corpora, with a focus on a few chosen relations.

First, we will take a brief overview of the datasets that we decided to use to complete the task. Then we will look at the transformer BERT models that we used for relation extraction.

Since our models were quite successfull at predicting these relations with predetermined entities, we also decided to give a try with implementing a custom model for automatic entity recognition for karstology domain.

Corpora and datasets

Besides the provided Karst corpus, we began with inspecting different corpora and datasets. We checked WikiData, NewYorkTimes (NYT) dataset and its subsets, NYT24 and NYT27. We also checked SloWnet and dataset form SemEval2010-Task8challenge. In the end, we decided to use the SemRelData dataset, because it has hypernym-hyponym pairs and is structurally similar to Karst. The linguists also provided us with the additional non-annotated karst corpus. Below we explore used datasets in more detail.

TermFrame - The Karst corpus

The TermFrame dataset contains relevant works in karstology in 3 languages – English, Croatian and Slovene. It includes books, reference works, PhD theses and scientific articles and other works related to Karst domain. It is available in plain text and chosen sentences were also manually annotated in WebAnno annotation tool [1] and are stored in .tsv file. Table 2 provides us with a brief overview of corpus by language. We chose to focus on English language since most datasets

are available in this language. There are four types of definition elements present in annotations, with the names and distribution visible on figure 1. Semantic relations and their distributions are visible on figure 2.

We also explored the distribution of words and phrases that are present in definition elements and semantic relations. For example, we prepared a small extract of words that are found in the annotated karst corpus. There is a small set of words that are highly likely to appear for certain types of relations. We later used this observation to derive certain rules for entity recognition.

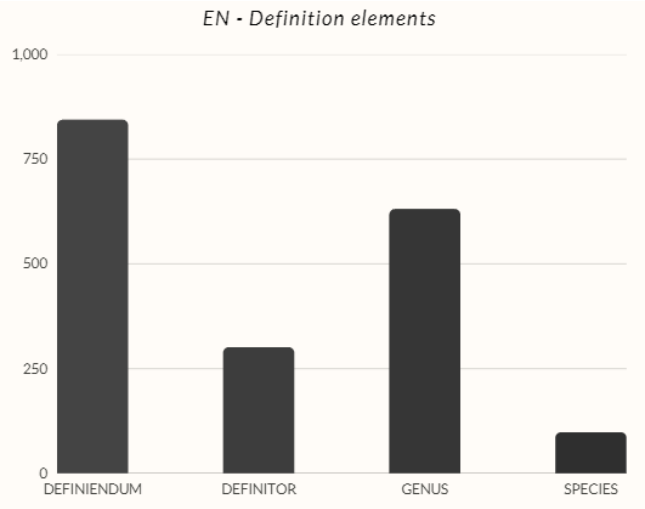

Figure 1. Number of definition elements by type for English in annotated definitions.

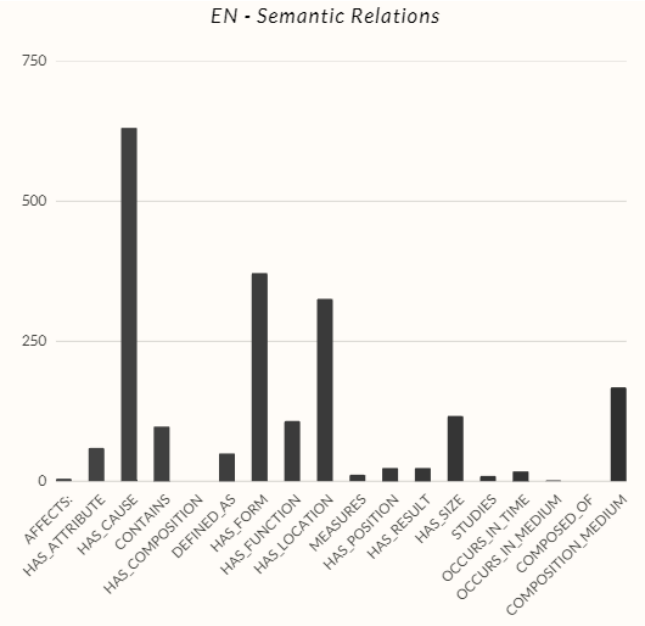

Figure 2. Number of semantic relations by relation type for English language in annotated definitions.

Table 1. Distribution of Semantic relation type for different languages (SemRelData dataset).

	English	German	Russian
Synonym	86	77	226
Co-Hyponym	281	336	296
Hypernym	553	508	296
Holonym	775	798	553

Table 2. Size of Karst 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

SemRelData dataset

SemRelData stands for Semantic Relation Dataset which is focused on contextual annotation of classical semantic relations and provides data for English, German and Russian language. It consists of news articles, encyclopedic articles and snippets from literary texts for all three languages and contains ≈ 60.000 tokens, with the distribution of semantic relation types shown in the table 1. The dataset is annotated in WebAnno and can be downloaded from University of Hamburg site [2]. More detailed analysis was done by Darina Benikova [3].

As we mentioned before, this dataset is very similar to the Karst dataset. It also has hypernym-hyponym pairs, which we can use to test and evaluate the models that were trained on Karst.

Non-annotated Karst corpus

The additional corpus was provided to us by the students from the Faculty of Arts. It was meant to evaluate the performance of our models. The non-annotated corpus and was given in plaintext, so we used two preprocessing methods, to prepare the data for our model. Firstly, we used Punkt Sentence Tokenizer [4] from Natural Language Toolkit, which divides a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences.

Although Punkt Sentence Tokenizer did manage to extract viable sentences from the whole corpus, there were still too many irrelevant sentences and cases which could not be used directly in BERT for relation classification. From this we extracted 46,604 sentences. Most of these sentences are not applicable for evaluation, so in the second step we filtered over the sentences. There were two steps of filtering, first we removed sentences that were too small or too large or had a specific structure (table of content, header, footers). Second, we used a collection of all words and phrases that occurred as difiniendums in the original Karst corpus. We then only kept the sentences that contain any of the difiniendum substrings, ex. “is a”, “are defined as”, “are made of”, etc. With this

method, we reduced the number of sentences to 12,143.

Linguist contribution

The main task in building an English corpus was collecting English texts on geomorphology, glaciology and geology that were, then, incorporated into the non-annotated Karst corpus created via Sketch Engine, ultimately consisting of 1,588,085 tokens.

Manual evaluation

Following the annotation process, the linguist wanted to analyze the predictions submitted by the model. This part proved to be quite burdensome, as the model was only able to indicate phrases containing a relation, but was unable to find it by itself. Another drawback of the model was the fact that it was not able to correctly separate the sentences, accounting for difficulties when attempting to locate full phrases. After examining the predictions, the linguist manually identified 20 sentences and annotated them according to hypernym-hyponym (genus-definiendum) relations. The team found that though the model had successfully recognized that all 20 sentences included hypernym-hyponym relations, it also listed many other sentences that did not contain such relations. Linguist annotation highlighted another obstacle: many of the manually annotated hypernyms and hyponyms were complex, containing more than one word, which likely confused the model when detecting definienda and genera. Ultimately, compared to manual annotations, model predictions proved to be insufficient and required thorough double-checking.

Automatic entity recognition

We tried to develop a tool for automatic entity recognition, which could extract entities that occur in a semantic relations. Firstly we used spaCy, which is existing tool for automatic entity recognition. Because of its poor performance on non-annotated Karst corpus, we also decided to try using custom rule based entity recognition.

spaCy

A spaCy model [5] was used before sentences were passing into BERT. The model did the automatic selection of entities which were then appropriately tagged.

The main problem of this approach was, that the model was recognizing entities that were often to long for any meaningful use in semantic relation extraction. This meant, that a lot of relations that were extracted from non-annotated corpus were either partly wrong or fully wrong.

Custom rules

We already mentioned in datasets section, that certain words appear more often in specific types of semantic relations. We tried to formulate a simple rules for determining if there is a semantic relationship or definition elements present in the sentence.

Let's take a look at *DEFINITOR* definition element for English. The top 5 definitors are:

1. Are (80 occurrences),
2. Is a (53 occurrences),
3. Is (18 occurrences),
4. Is defined as (11 occurrences),
5. The term (11 occurrences).

Together this sums up to 173 of 300 definitors. This means that by using just top 5 definitors, we can already cover more than 50% of all definitors. This does not apply for other definition elements such as definiendum, which are mostly unique phrases. Using top 5 definiendums only covers around 4% of all definiendums.

Another problem is trying to figure out the correct order of entities. Is the definiendum located after or before the genus or species? There is often a case, that lots of words are hypernym and hyponym at the same time in relation to some other words. Therefore it is not trivial to guess which one is correct. Though we can assume that there is a high probability that relation is present in the sentence if certain words are matched. This however is often not enough to correctly predict the presence of relations and find entities.

BERT

ALBERT

Results

Conclusion

Relationship extraction models performed as expected and did, in some cases, even surpased our expectation. Even with a short fine-tuning time, models were able to learn quickly and they can also be used interchangable between two chosen datasets. We compared the performance of two transformer models, BERT and ALBERT and concluded, that ALBERT is more fit to the task of relationship extraction.

However we have to point out, that we were largely unsuccessful in achieving a good automatic entity recognition. It seems, that because of the Karst domain specifics, it is not as easy to suggest an entities since they can be very specific and vary a lot.

In general we accomplished the basic tasks. There is also a lot of space for improvement. With the knowledge we acquired during working on this project, we would definitely approach some things differently, such as the construction of a general pipeline for preparing training and testing data using different datasets.

References

[1] WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. Available at: https://webanno.github.io/webanno/. [Accessed: 20. 5. 2022].

[2] SemRelDataset. Available at: https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/semreldata.html. [Accessed: 20. 4. 2022].

[3] Darina Benikova and Chris Biemann. Semreldata — multilingual contextual annotation of semantic relations between nominals: Dataset and guidelines. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[4] Punkt Sentance Tokenizer. Available at: https://www.nltk.org/_modules/nltk/tokenize/punkt.html. [Accessed: 24. 5. 2022].

[5] Spacy for entity Recognition. Available at: https://spacy.io/usage/linguistic-features. [Accessed: 20. 5. 2022].