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

Natural Language Processing - Report for 2nd submission (Task 2)

Lara Kuhelj, Gregor Novak, Jan Pelicon, Tomaž Štrus

Abstract

For the task of hypernym-hyponym extraction and extraction of new instances of semantic relations from domain-specific corpus we analyzed several different datasets with different types of semantic relationships and domains. For now, we fine-tuned large BERT model on SemRelData relational dataset and started experimenting with training BERT from scratch. In the future we want to test different scenarios, languages and datasets in order to choose the best setting and also improve the current model for the extraction of hypernym-hyponym pairs. Additionaly, we will start work on the remaining parts of the task.

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 these 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 of semantic relations from provided Karst corpus and other explored datasets and corpuses, with a focus on a few chosen relations.

Corpuses and datasets

Besides the provided Karst corpus, we explored additional corpuses and datasets that capture semantic relationships. They can be used as additional training or evaluation material.

The Karst corpus

The corpus contains relevant works in karstology. This includes books, reference works, PhD theses and scientific articles. It is provided as in plain text and in the ClassLa tagged format. Annotated definitions and semantic relations (15 types) are also provided in WebAnno format. From the paper, we can see the size of the corpus in the table 1. We will manually parse the prepared content into a suitable form for the pretraining, fine-tuning or testing the used models.

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

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 2. The dataset is annotated in WebAnno and can be downloaded from University of Hamburg site [1]. More detailed analysis was done by Darina Benikova [2].

WordNet - SloWnet

The original WordNet [3][4] is an English lexical database where nouns, verbs, adjectives and adverbs are grouped into sets of synonyms called synsets which are linked by semantic

Table 2. 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

and lexical relations. The most frequent relation in synsets is hyperonym-hyponym also called ISA relation. We explored SloWnet, a Slovenian WordNet, from Open Multilingual WordNet [5] which contains of 42,583 synsets, 40,233 words and 70,947 senses. It is available in XML or simple tab format.

NewYorkTimes and WikiData

The New York Times [6] Annotated Corpus consists of articles written by the New York Times between 1987 and 2007. We explored the main NYT dataset and the two derivations of a New York Time dataset called NYT24 and NYT27 [7][8], where the number specifies how many types of relations the dataset contains. Datasets relations are expressed using relation hierarchy, where four main relations categories are: people, business, person, location. The relation information is stored as a JSON structure with instances, aliases, description and label for each entity with relations to other entities. A similar dataset is WikiData [9], which is a collection of texts from free knowledge base.

SemEval-2010 Task 8

SemEval-2010 Task 8 [10] is a dataset for multi-way classification of mutually exclusive semantic relations between pairs of nominals. The dataset was used for a competition in order to compare different models and approaches to the problem. It includes 10 different classes: Other, Cause-Effect, Product-Producer, Entity-Origin, Instrument-Agency, Component-Whole, Content-Container, Entity-Destination, Member-Collection, Message-Topic. The entities that are present in semantic relations are marked using XML tags, where a type of relation is defined between two entities.

Implementation of relation extraction

First, we had to determine which datasets to use for a specific task. For extraction of hypernym-hyponym pairs from a domain-specific corpus, we decided to use the annotated Karst corpus for testing the model and SemRelData dataset for fine-tuning since they are both annotated using the same WebAnno annotation tool and contain hypernym-hyponym relations.

We used a PyTorch implementation of BERT from 2019 paper "Matching the Blanks: Distributional Similarity for Relation Learning" [11] which has provided an improvement over previous methods on SemEval 2010 Task 8.

BERT model tries to predict a relation $r = (x, e_1, e_2)$ for sequence of tokens x and two marked entities e_1, e_2.

First we tried using the mentioned BERT model without pretraining step, fine-tuning it on English version of SemRelData dataset. SemRelData relations are composed only from one word. This indicates that phrase with tag e_2 is in this relation with phrase with tag e_1. For example, for Hypernym relation, phrase with tag e_2 is the hypernym of phrase with tag e_1.

Because some sequences in SemRelData were longer than 512 characters (consisting of multiple sentences), which is the maximum supported sequence length in the used BERT implementation, we tried to shorten some sequences with removing the sentences that don't include either of the e_1 and e_2 tags. Some sequences were still too long after this procedure and were thus ignored. In the future, we plan to experiment with more complex ways of extracting only the relevant sequence parts around e_1 and e_2 tags in sequences, so that we can use as many sequences in the original dataset as possible. With the current approach we gained 810 different relations which we used for fine-tuning. Their distributions are shown in table 3.

Table 3. Distribution of relation types from SemRelData dataset, used for fine-tuning the BERT model.

	Number of relations
Synonym	39
Co-Hyponym	140
Hypernym	295
Holonym	336
Total	810

In Karst corpus, we tagged e_1 and e_2 by hand with the knowledge from provided annotated data. We used e_1 tag for definiendum phrases (hyponyms) and e_2 tags for genus phrases.

We fine-tuned the BERT model without pretraining step. We used 11 epochs for fine-tuning. Last epoch achieved roughly 0.80 accuracy on training data, loss value was around 0.50.

Results

After fine-tuning was completed, we tested the resulting model on the annotated Karst corpus with hyponym-hypernym relations. As mentioned before, we always tagged definiendum phrases with e_1 tags and genus phrases with e_2 tags, indicating that genus phrase is a hypernym of definiendum phrase. Some annotated sequences are missing the genus words. Because of this, not all the annotated sequences in Karst corpus could be tagged with e_1 and e_2 tags. We did not perform relation inferring on these sequences, since we are missing both words/phrases in the relation. There were 127 occurrences of such sequences. If a sequence contained more than two tagged phrases (generally more than one genus phrase), it was

inferred only once. In the future, we will also try separating such tag pairs into multiple tagged instances of the same sequence, each containing only one e_1 and one e_2 tag.

Distribution of detected relations are shown in table 4. Percentage values are rounded to two decimals.

Table 4. Results achieved on annotated Karst corpus by BERT model, fine-tuned with SemRelData dataset.

	Detected relations	Percentage
Hypernym	414	72.76 %
Holonym	132	23.20 %
Co-Hyponym	15	2.64 %
Synonym	8	1.41 %
Total	569	100 %

The most common predicted relation was Hypernym, followed by Holonym. If we only consider Hypernym relations as True Positive, model achieved 72.76 % accuracy. But it could also be argued that Holonym relations are to some extent semantically similar to hypernym relations (For example, word "Face" is holonym of the word "Eye", indicating a hierarchical relation).

At the first glance, these results look promising, especially if we consider that we used basic approaches without much modification. But on the other hand, we can also see that Hypernym and Holonym are the two most frequent relations appearing in the sequences, used for fine-tuning. Because of this, the model could perhaps detect them more frequently because it knows them the best. To gain better insight into the quality of the model, we will need to test it on some other dataset, for which we know that it contains relations, which are semantically really different from hypernym-hyponym so that we can consider them as negative examples.

Future work

Currently, we have preliminary model for extraction of hypernym-hyponym pairs from the specific corpus, based on the BERT model. We used BERT model without pretraining step. Because of this, we plan to create another version of the model, where we will also use pretraining. Additionally, we plan to experiment with different datasets for training, as well as methods, other than BERT, which we mentioned in the previous submission.

We still need to implement a solution, which will be able to automatically suggest the two words/phrases for relation extraction. Currently, we determined such phrases in Karst corpus by hand, considering the provided annotated data.

We will also start working on the second part of the specified task, i.e., relation extraction (other than hyponym-hypernym), from domain specific corpus. Here we plan to use similar approaches as in the first part of the task and modify them for the specific relations.

When we are finished with semantic relation extraction, we will work together with student from Faculty of Arts to perform qualitative analysis on the extracted data, trying to determine the sequence properties where models perform well and sequence properties where models fail to achieve good results.

Conclusion

Our plan is to experiment with available datasets as much as possible in order to come with multiple working and successful models for hypernym-hyponym pair extraction, as well as other specific relation extraction for the second part of the task. We will first implement the tools using English BERT model and datasets and will later try to expand this with different language. Furthermore, we will continue the work on second task of extracting new instances of relations from domain specific corpus and experiment with methods, other than BERT. If possible, we will also try to combine results of different methods to achieve better accuracy. After the final version of relation extraction, we will work together with student from Faculty of Arts to analyze the extracted data.

References

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

[2] 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).

[3] George A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995.

[4] Dagobert Soergel. Wordnet. an electronic lexical database. 10 1998.

[5] OpenMultilingualWordnet. Available at: http://compling.hss.ntu.edu.sg/omw/. [Accessed: 20. 4. 2022].

[6] Evan Sandhaus. Nyt - new york times annotated corpus. the new york times annotated corpus ldc2008t19, 2008.

[7] Sebastian Riedel, Limin Yao, and Andrew Mccallum. Modeling relations and their mentions without labeled text. pages 148–163, 09 2010.

[8] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel Weld. Knowledge-based weak supervision for information extraction of overlapping relations. volume 1, pages 541–550, 01 2011.

[9] Wikidata. Available at: https://www.wikidata.org/wiki/Wikidata:Main_Page. [Accessed: 20. 4. 2022].

[10] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In **SEMEVAL*, 2010.

[11] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning, 06 2019.