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

Automatic semantic relation extraction

Aljoša Koren, Klemen Škrlj, and Tilen Kavčič

Abstract

Keywords
Semantic relation extraction, TermFrame knowledge base

Advisors: Slavko Žitnik

Introduction

The purpose of our project is to research different approaches to automatic semantic relation extraction. Firstly, we will train and evaluate our models on the TermFrame knowledge base [1] which is a Karstology domain specific corpora for three languages; Slovenian, English and Croatian. The corpora contains scientific texts from the field of karstology. Combined the corpora contains more than 5 million tokens from 160 documents as it is shown in Table 1. TermFrame corpus

	English	Slovene	Croatian
Tokens	2,721,042	1,208,240	1,229,368
Words	2,195,982	987,801	969,735
Sentences	97,187	51,990	53,017
Documents	57	60	43

Table 1. The TermFrame corpora information.

is annotated with 5 informations; canonical form, semantic category, definition element, semantic relation and relation definitor. We will be focusing on the definition element as it marks the definiendum and the genus. Definiendum is the term which is defined in the definition sentence, while genus represents its hypernym. We will then try to define other relationships, not just the hyponym-hypernym relations. For this we will use the semantic relation information, as it gives information about properties or features of the term that is being defined, such as location, cause, size etc.

Corpus

- https://termframe.ff.uni-lj.si/visualizations/
- http://compling.hss.ntu.edu.sg/omw/

Related works

There are usually two types of algorithms for discovering the hyponym-hypernym relation; pattern-based and distributional methods. The pattern-based approach is usually time-consuming and language dependent, even if we take the same language from a different time period. Distributional methods can be supervised or unsupervised. They use word distribution to extract hypernyms. Roller et al. [2] compared both approaches in 2018. Both approaches use co-occurrences within a context, however pattern-based use predefined manually chosen patterns, while distributional methods use unconstrained word co-occurrences. They have extracted simple Hearst patterns and also broader patterns, took frequency of occurrences and sparsity into account and postpreprocessing which removed pairs that did not occur in enough sentences. This method was discovered to be better then the rest distributional methods they were comparing it with. The work done by Atzori and Balloccu [3] used a unsupervised learning for hypernym discovery. They used cosine distance in vector word embeddings as it was done before, but they added rank weighted by word frequencies in a corpus and level of similarity, to remove the semantic relations that might not be in the hyponym-hypernym relation. For example their system allows the inputs of co-hyponymes to discover a common hypernym. This is especially useful when one word could mean two different things. In the paper they show example for ”apple”. If it is the only input, it might output that its hypernym is ”company”, yet if we add a co-hyponym ”pear” it will defiantly output ”fruit”. Their system is domain and language independent, because they do not use Hearst patterns [4] (realtions of the form x is-a y) or stopwords, but solely unstructured data. Their algoritm firstly processes the text by removing punctuations, making all words lowercased etc. They use word2vec algorithm which learns word associ-

ations from corpus. Then in the second step they extract the potential hypernyms by Nearest Neighbours Search with cosine similarity. They can use a single labeling or set labeling if we have multiple words with the same hypernym. At the end the algorithm gives scores to the returned candidates and returns the best guesses.

Paper [5] describes results of a challenge where competitors had to extract hypernym-hyponym relations between a given list of domain-specific terms. The best approach was based on Hearst patterns where the team made use of large web corpus. The distributional approaches got competitive recall but struggled with precision. In a competition [6] teams were presented with a task of finding suitable hypernyms of a given word from the target corpus. The corpuses were multilingual and also domain-specific. Here a system that learned embeddings of hyponym-hypernym pairs combined with unsupervised learning of Hearst-style patterns performed the best. Overall supervised methods showed clear superiority over unsupervised ones.

Dependency trees are commonly used to tackle extraction of new intra-sentence instances of semantic relations. To keep only the most relevant information pruning is used. Zhijiang Guo , Yan Zhang and Wei Lu (2019) [7], for example, introduce Attention Guided Graph Convolutional Networks s (AGGCNs) which use a soft-pruning approach to clean up the dependency tree and keep the relevant sub-structures useful for the relation extraction task. AGGCNs transform the dependency tree into a fully connected edge-weighted graph. Weights represent the relatedness between nodes. Since dependency trees don't capture inter-sentence relations, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are often used. Sahu et al. 2019 [8] employ their edge Graph CNN (GCNN) model to capture local and non-local dependencies. The graph nodes represent words while edges correspond to semantic dependencies. They use MIL-based bi-affine pairwise scoring function to infer relations between entities from the entity node representations.

Methods

Results

Discussion

Acknowledgments

References

[1] Špela Vintar, Amanda Saksida, Katarina Vrtovec, and Uroš Stepišnik. Modelling specialized knowledge with conceptual frames: The termframe approach to a structured visual domain representation. 2019.

[2] Stephen Roller, Douwe Kiela, and Maximilian Nickel. Hearst patterns revisited: Automatic hypernym detection from large text corpora. *CoRR*, abs/1806.03191, 2018.

[3] Maurizio Atzori and Simone Balloccu. Fully-unsupervised embeddings-based hypernym discovery. *Information*, 11(5):268, May 2020.

[4] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, 1992.

[5] Georgeta Bordea, Els Lefever, and Paul Buitelaar. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California, June 2016. Association for Computational Linguistics.

[6] Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[7] Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction, 03 2022.

[8] Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Inter-sentence relation extraction with document-level graph convolutional neural network, 03 2022.