Apollo at SemEval-2018 Task 9: Detecting Hypernymy Relations Using Syntactic Dependencies

Mihaela Onofrei, Ionuț Hulub,
Diana Trandabăț, Daniela Gîfu

University Alexandru Ioan Cuza of Iași, Romania
Institute of Computer Science of the Romanian Academy, Iași Branch
Cognos Business Consulting S.R.L., 32 Bd. Regina Maria, Bucharest, Romania
{mihaela.onofrei, ionut.hulub, daniela.gifu, dtrandabat} @info.uaic.ro

Abstract

This paper presents the participation of Apollo's team in the SemEval-2018 Task 9 "Hypernym Discovery", Subtask 1: "General-Purpose Hypernym Discovery", which tries to produce a ranked list of hypernyms for a specific term. We propose a novel approach for automatic extraction of hypernymy relations from a corpus by using dependency patterns. The results show that the application of these patterns leads to a higher score than using the traditional lexical patterns.

Keywords: hypernymy relations, semantic relations, corpus, taxonomy, syntactic dependencies.

1 Introduction

This paper presents the Apollo team's system for hypernym discovery which participated in task 9 of Semeval 2018 (Camacho-Collados et al., 2018) based on unsupervised machine learning. It is a rule-based system that exploits syntactic dependency paths that generalize Hearst-style lexical patterns.

The paper is structured in 4 sections: this section presents existing approaches for automatic extraction of hypernymy relations, Section 2 contains the current system architecture. The next section presents the web interface of the project, and, finally, Section 4 briefly analyses the results and drafts some conclusions.

Since language is a "vital organ", constantly evolving and changing over time, there are many words which lose one of their meanings or attach a new meaning. For instance, when searching the word "apple" in WordNet (Miller, 1995), it appears defined as "fruit with red or yellow or green skin and sweet to tart crisp whitish flesh" and "native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits" but searching in British National Corpus[1], we will remark that the term is used more frequently as a named entity (referring to a "company").

From this point of view, we consider that developing a system for hypernym discovery that uses linguistic features from a corpus could be more useful for this task than using a manually-crafted taxonomy.

It is well known that in natural language processing (NLP), one of the biggest challenges is to understand the meaning of words. Also, detecting hypernymy relations is an important task in NLP, which has been pursued for over two decades, and it is addressed in the literature using two complementary approaches: rule-based and distributional

[1] https://corpus.byu.edu/bnc/

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 898–902
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

methods. Rule-based methods (Hearst, 1992; Snow et al., 2004) base the decision on the lexico-syntactic paths connecting the joint occurrences of two or more terms in a corpus. In the case of supervised distributional methods (Baroni et al., 2012; Roller et al., 2014, Weeds et al., 2014; Levy et al., 2015, Kruszewski et al., 2015), term-pair is represented using some combination of the terms' embedding vectors.

This challenge has been shown to directly help in downstream applications such automatic hypernymy detection is useful for NLP tasks such as: taxonomy creation, recognizing textual entailment, text generation, Question Answering systems, semantic search, Natural Language Inference, Coreference Resolution and many others.

Traditional procedures to evaluate taxonomies have focused on measuring the quality of the edges, i.e., assessing the quality of the *is-a* relations. This process typically consists of extracting a random sample of edges and manually labeling them by human judges. In addition to the manual effort required to perform this evaluation, this procedure is not easily replicable from taxonomy to taxonomy (which would most likely include different sets of concepts), and do not reflect the overall quality of a taxonomy. Moreover, some taxonomy learning approaches link their concepts to existing resources such as Wikipedia.

2 A new Approach to Detect Hypernymy Relation

The main purpose of this project was to identify the best (set of) candidate hypernyms for a certain term from the given corpus[2].

In our system, we considered the rule-based approach and, in order to extract the corresponding patterns, we used syntactic dependencies relations (Universal Dependencies Parser[3]).

Below, we present our method of extracting hypernyms from text:

- **Tokenization**: sentence boundaries are detected and punctuation signs are separated from words;

- **Part-of-speech tagging**: the process of assigning a part-of-speech or lexical class marker to each word in a corpus. Words in natural languages usually encode many pieces of information, such as: *what the word "means" in the real world, what categories, if any, the word belongs to, what is the function of the word in the sentence?* Many language processing applications need to extract the information encoded in the words. Parsers which analyze sentence structure need to know/check agreement between: subjects and verbs, adjectives and nouns, determiners and nouns, etc. Information retrieval systems benefit from know what the stem of a word is. Machine translation systems need to analyze words to their components and generate words with specific features in the target language.

- **Dependency parsing:** the syntactic parsing of a sentence consists of finding the correct syntactic structure of that sentence in a given formalism/grammar. Dependency parsing structure consists of lexical items, linked by binary asymmetric relations called dependencies. It is interested in grammatical relations between individual words (governing & dependent words), it does not propose a recursive structure, rather a network of relations. These relations can also have labels and the phrasal nodes are missing in the dependency structure, when compared to constituency structure.

One of the boosts for this approach was to develop new dependency patterns for identifying hypernymy relations from text that are based on dependency relations. The increased popularity and the universal inventory of categories and guidelines (which facilitate annotation across languages) of Universal Dependencies determined us to use this resource in order to automatically extract the hypernyms from the corpus.

[2] For this subtask, we used the 3-billion-word UMBC corpus, which consists of paragraphs extracted from the web as part of the Stanford WebBase Project. This is a very large corpus containing information from different domains.

[3] Universal Dependencies (UD) is a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing more than 100 treebanks in over 60 languages.

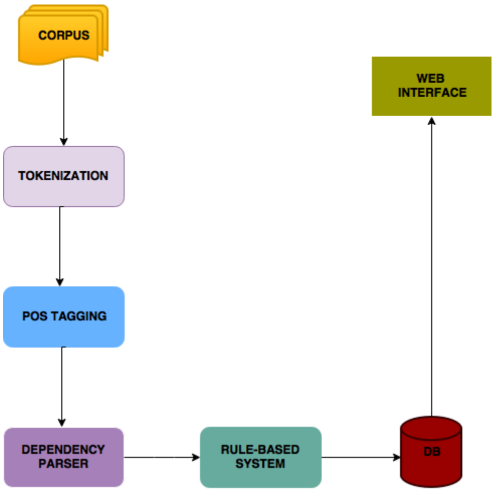


Figure1: Project's architecture

In this manner, we managed to compress a list of 44 lexico-syntactic patterns used for the hypernyms extraction[4] in only 8 dependencies patterns. In the next lines, we present few examples of lexico-syntactic patterns that were replaced by dependencies patterns:

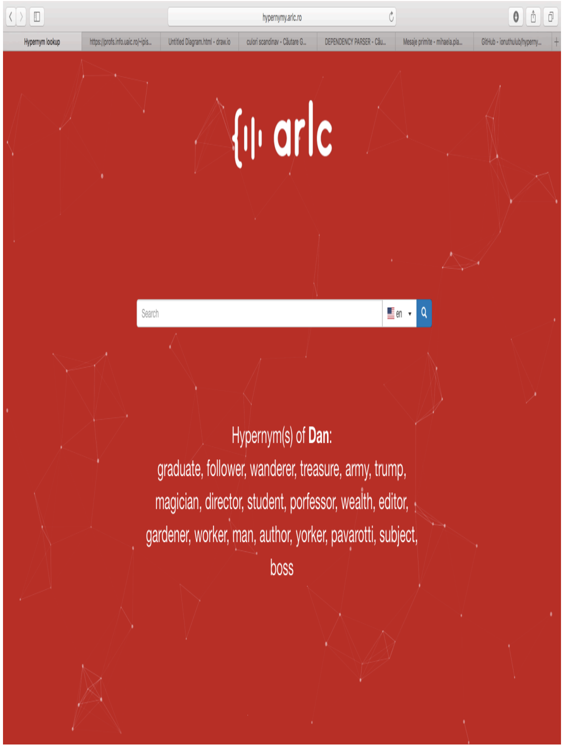
{X and other Y; X or other Y; X and any other Y; X and some other Y; Y other than X; X like other Y; Y other than X} → replaced by ***X "amod" Y;***

{X is a Y; X was a Y; X are a Y; X are Y; X will be a Y; X is an adj Y; X was a adj. Y; X are a adj. Y; X was a adj. Y; X are examples of Y; X is example of Y; Y for example X; examples of Y is X; examples of Y are X; X which is named Y; X which is called Y; Y which are similar to X; Y which is similar to X} replaced by **X "nmod" Y.**

Because we used syntactic dependencies relations (no lexical patterns were involved), our system is language independent. Unfortunately, the limited hardware resources determined us to run our system only in English but we are looking forward to running it in both Spanish and Italian.

3 The Web Interface

The interface[5] was implemented in the form of a website. The site is backed by a Mongodb database. When a user types in a query and hits enter a post request is sent and the backend will do some processing on the query (tokenizing, lemmatizing) and then search in the database. The results are then sent back to the user where they are rendered.



Hypernym(s) of **Dan**:
graduate, follower, wanderer, treasure, army, trump, magician, director, student, porfessor, wealth, editor, gardener, worker, man, author, yorker, pavarotti, subject, boss

Figure 2: Project's interface

4 Results

We consider that a qualitative way of analyzing our system is to look at which relations are more productive. Table 1 presents the percentages of the most representative syntactic relations which we have identified. While some relations have not been very fruitful (such as ***X "obj" Y***, for insance), others, instead, have been very productive, generating tens of thousands relations.

[4] http://webdatacommons.org/isadb/lrec2016.pdf

[5] http://hypernymy.arlc.ro

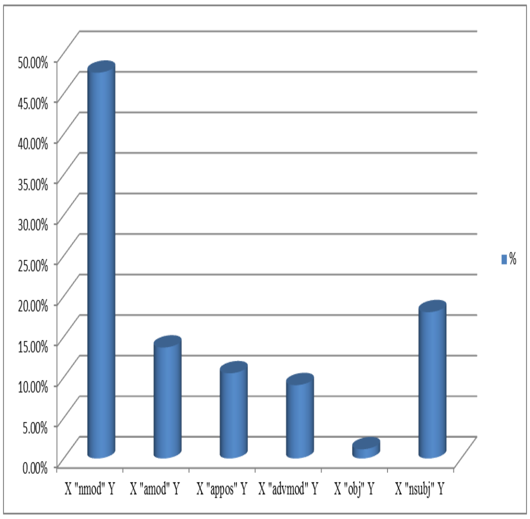


Table 1: Percentages of the identified syntactic relations

The project's results show that we have managed to accomplish the main objective of this project, to outperform the random strategy. The lower scores have been obtained for multiword expressions, for which we plan to add dedicated modules.

An issue that we have noticed was that the given vocabulary was quite restrictive, for instance, it contains words like "above-water", "artesian water", "bath water" etc., but it doesn't contain the word "water" (we had a case when our system identified the word "water" as a hypernym and it was a correct hypernym, but due to the fact that the vocabulary doesn't contain the word "water", it cannot be evaluated) and many other examples like this.

Acknowledgements

This survey was published with the support by two grants of the Romanian National Authority for Scientific Research and Innovation, UEFISCDI, project number PN-III-P2-2.1-BG-2016-0390, contract 126BG/2016 and project number PN-III-P1-1.2-PCCDI-2017-0818, contract 73PCCDI/2018 within PNCDI III, and partially by the README project "Interactive and Innovative application for evaluating the readability of texts in Romanian Language and for improving users' writing styles", contract no. 114/15.09.2017, MySMIS 2014 code 119286.

References

Baroni, M., Lenci, A. 2011. *How we blessed distributional semantic evaluation.* In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, pages 1-10.

Bordea, G., Buitelaar, P., Faralli, S., Navigli, R. 2015. *Semeval-2015 task 17: Taxonomy extraction evaluation (Texeval).* In *Proceedings of the SemEval workshop.*

Camacho-Collados, J. 2017. Why we have switched from building full-fledged taxonomies to simply detecting hypernymy relations. *arXiv preprint arXiv:1703.04178.*

Camacho-Collados, J., Deli Bovi, C., Espinosa-Anke, L., Oramas, S., Pasini, T., Santus, E., Scwartz, V., Navigli, R., Saggion, H. 2018. *SemEval-2018 Task 9: Hypernymy Discovery.* In Proceedings of the 12th International Workshop on Semantic Evaluation (Sem-Eval2018), New Orleans, LA, United States. Association for Computational Linguistics.

Hearst, M. 1992. *Automatic acquisition of hyponyms from large text corpora.* In *ACL*, pages 539-545.

Kruszewski, G., Paperno, D., Baroni, M. 2015. *Deriving Boolean structures from distributional vectors. Transactions of the Association for Computational Linguistics*, 3:375-388.

Levy, O., Remus, S., Biemann, C., Dagan, I. Ramat-Gan, I. 2015. Do supervised distributional methods really learn lexical inference relations? *In Proceedings of NAACL, pages 970–976.*

Miller, G. 1995. *WordNet: A lexical database for English. Communications of the ACM,* 38(11): 39-41.

Roller, S., Erk, K. 2016. *Relations such as Hypernymy: Identifying and Exploiting Hearst Patterns in Distributional Vectors for Lexical Entailment.* In *Proceedings of EMNLP*, pages 2163–2172.

Roller, S., Erk, K., Boleda, G. 2014. *Inclusive yet selective: Supervised distributional hypernymy detection.* In *COLING*, pages 1025-1036.

Shwartz, V., Santus, E., Schlechtweg, D. 2017. *Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection.* In *Proceedings of EACL,* pages 65–75.

Snow, R., Jurafsky, D., Y Ng, A. 2004. *Learning syntactic patterns for automatic hypernym discovery.* In *NIPS.*

Wang, Ch., He, X., Zho, A. 2017. *A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances.* In *Proceedings of EMNLP,* pages 1201–1214.

Weeds, J., Clarke, D., Reffin, J., Weir, D., Keller, B. 2014. *Learning to distinguish hypernyms and co-hyponyms.* In *COLING*, pages 2249-2259.