

Extended Abstract: Extracting Semantic Relations of High Textual Complexity

M. Döhring^{1,2}, N. Reuschling^{1,2}, U. Störl¹, A. Berlea²

doehring@stud.fbi.h-da.de, reuschling@stud.fbi.h-da.de,
u.stoerl@fbi.h-da.de, alexandru.berlea@sap.com

¹ Hochschule Darmstadt, ² SAP Research

1 MOTIVATION AND SCOPE

The ability to automatically gain information from unstructured data is an essential task for the purpose of exploiting the enormous potential of information available on the Web. In particular, the promise of the *Semantic Web*, more specifically of being able to automatically capture the meaning of Web content, requires the ability to extract semantic relations from unstructured text. Semantic Relation Extraction builds upon technologies from Natural Language Processing (NLP). While NLP techniques have greatly matured throughout the last decade, improvements are still needed in order to be able to use it fully automatically. For example, despite significant efforts invested in relation extraction in research areas such as Ontology Learning and Ontology Population [BCM05], a qualitatively acceptable procedure for the automatic extraction of ontologies is still a long way off [Zho07]. A main disruptive factor is the yet insufficient addressing of relations with high textual complexity in terms of nested or interchanged syntactic structures as well as the presence of multiple named entities in a sentence. With the biomedical domain in focus, [JVD07] points out the general relevance of sentence complexity for data misinterpretation. [KKT03], applying most prevalent NLP techniques for extracting relations in the biomedical domain, state that 40% of their false negatives are due to the complexity of sentences. Motivated by current research presented in [Doe08], we manually validated similar figures also for the task of recognizing company cooperations in business texts detected by OpenCalais [Reu], a publicly available system for named entity recognition and relation extraction. Below, we give an authentic example for such a sentence with a high degree of complexity.

To provide another authorized Web outlet for its assets, *Viacom* signed a licensing agreement last month with *YouTube* rival *Joost* to distribute video content from *Comedy Central* and *MTV* over the Internet.

What are the main obstacles when automatically extracting relations from the above sentence? Currently, one limitation of prevalent approaches for relation extraction arises from

the expressiveness of the patterns that they are able to recognize. Most approaches use *shallow* pattern matching (in some way similar to regular expressions), that is they are partly oblivious of the grammatical structure of text sentences. In the above example, they could wrongly detect a collaboration example between *Viacom* and *YouTube*, negatively influencing precision as well as recall. The shallow patterns are intrinsically dependent on the order of words in sentences and fail to capture a relation if the relative order of the words changes, even though the expressed relation remains the same. Uninvolved named entities even complicate the detection of a relation or the correct assignment entities directly involved in the relation. We therefore require approaches which abstract away from the representation of a sentence as a sequence of tokens to representations that are more stable with respect to the relative order of words as long as the relation stated among the words or entities of interest remains the same.

One component for the realization of this idea is the exploitation of grammatical dependencies, i.e. linguistic structures that go beyond the flat sentence level, for relation extraction. Scientific work in this area has increased recently, although most of them either rely on a trained model or only address narrow types of relations (such as is-a or part-of). We show that it is possible to achieve improvements in accuracy for relations of high textual complexity also with a defined unsupervised procedure.

2 OUTLINE OF SOLUTION APPROACH

In order to cope with the challenge of extracting high textual complexity semantic relations, we elaborated a generalizable hybrid procedure consisting of shallow pattern matching over word resp. part-of-speech (POS) tokens and the exploitation of grammatical dependencies on the sentence level. With the help of a fixed set of linguistic rules on the tree level, a parse tree with hierarchical sentence chunks and POS-tagged words can be mapped to a dependency tree with each word assigned to its governor [DMM06]. For instance, it indicates the dependency of adjectives on the nouns they modify or the dependency of nouns on their corresponding verbs. In Figure 1, we use a combined representation of the parse tree together with the dependency tree. The dotted arcs are labeled with the dependency type and concurrently indicate the direction of a dependency by pointing on the governing word. For example, the noun '*butterfly*' depends on the verb '*to catch*' via a direct object relation. Basically, our approach relies on the shortest path hypothesis of [BM05], which states that the information needed to determine a semantic relation between two entities in a text lies on the shortest path among them following their grammatical dependency relations. However, we discovered the need to extend this basic assumption to properly work with a current state-of-the-art sentence parser from [KM03], since it sometimes fails to recognize the correct grammatical dependencies. We therefore defined a set of pre- and postprocessing rules in addition to our matching algorithm handling special sentence constellations as for example conjunctions or the above mentioned multiple entity occurrences.

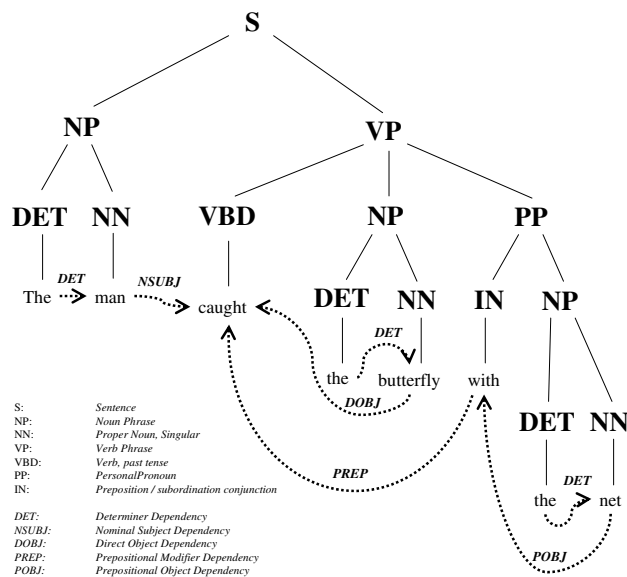


Abbildung 1: Example of a deep parse tree with dependency annotations.

The main advantage of the approach is its ability to overcome the interfering impact of „intermediate“ sentence elements between two candidate entities for a semantic relation in a very long or complex sentence. In our initial example sentence, the direct object grammatical dependency of *Joost* to *Viacom* helps detecting the correct semantic relation.

3 RESULTS AND FUTURE WORK

To be able to assess the usefulness of the proposed procedure especially for our particular application domain of high sentence text complexity, we defined a novel measure for determining the *complexity* of a semantic relation on the textual level. This metric was used for the practical application of our approach on a large¹ up-to-date dataset consisting of business news articles. Our evaluation was performed in direct comparison to OpenCalais with regard to the detection of collaboration relations between companies in the corpus. We chose the top-150 sentences according to the complexity metric and manually evaluated the correctness of the presence or absence of discovered semantic relations. According to the chi-square test value with Yates correction for two categories of data (correctly and wrongly detected relations in the sentences), our presented procedure significantly outperforms the results of OpenCalais at a confidence level of 99,9%. We can conclude that our presented procedure bears potential for improving relation extraction accuracy especially

¹550.000 candidate sentences

for relations of high textual complexity.

Some issues remain subject to further research. Testing against a manually built-up gold standard (consisting of about 200 randomly chosen sentences from the corpus), both systems yielded relatively low overall precision and recall values (about 35%), partly due to 'hidden' semantic relations which can only be discovered by a human reader with the use of context and/or world knowledge. A second issue is that in sentences of low complexity according to our metric, OpenCalais seems to yield better results. Consequently, it has to be examined how established procedures for semantic relation extraction in *simple* text corpora can be combined with our approach. Furthermore, a part of the rules and procedures employed by our system were only introduced due to errors and inaccuracies of the employed sentence parser and thus do not follow any original linguistic motivation. That means, the improvement of sentence deep parsing performance or the employment of a different parsing technique may change circumstances and conditions for our system and therefore might require its adjustment. Additionally, the exploitation of automatic pattern learning algorithms has not been addressed yet.

Literatur

- [BCM05] P. Buitelaar, P. Cimiano und B. Magnini. Ontology Learning from Text: An Overview. In P-Buitelaar, P. Cimiano und B. Magnini, Hrsg., *Ontology Learning from Text: Methods, Evaluation and Applications*, Jgg. 123 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2005.
- [BM05] Razvan C. Bunescu und Raymond J. Mooney. A shortest path dependency kernel for relation extraction. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Seiten 724–731, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
- [DMM06] M.-C. De Marneffe, B. Maccartney und C. D. Manning. Generating typed dependency parses from phrase structure parses. In *In LREC 2006*, Seiten 449–454, 2006.
- [Doe08] Markus Doehring. EXTRACONN: Extraction and Analysis of Company Networks from News. In Rainer Ruggaber, Hrsg., *Proceedings of the 1st Internet of Services Doctoral Symposium 2008 at International Conference on Interoperability of Enterprise Systems and Applications (I-ESA'08)*, Berlin, Germany, March 25th 2008.
- [JVD07] H. Jose, T. Vadivukarasi und J. Devakumar. Extraction of protein interaction data: a comparative analysis of methods in use. *EURASIP J. Bioinformatics Syst. Biol.*, 7(4):1–9, 2007.
- [KKT03] A. Koike, Y. Kobayashi und T. Takagi. Kinase pathway database: An integrated protein-kinase and NLP-based protein-interaction resource. *Genome Research*, 13:1231–1243, 2003.
- [KM03] D. Klein und C. D. Manning. Accurate unlexicalized parsing. In *ACL-41*, Seiten 423–430, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
- [Reu] Reuters. OpenCalais - Calais WebService. <http://www.openalais.com>.
- [Zho07] Lina Zhou. Ontology learning: state of the art and open issues. *Inf. Technol. and Management*, 8(3):241–252, 2007.