Ontology-based disambiguation of the semantic relation between the heads of two noun phrases

Thomas Vestskov Terney and **Tine Lassen**
Roskilde University, Department of Computer Science
tvt@ruc.dk, tlassen@ruc.dk,

Abstract

In order to facilitate content based information retrieval we need methods for analyzing the semantic structures in the text. In order to facilitate this, we propose a straightforward method for identifying semantic relations between two concepts in an ontology based on supervised machine learning. More specifically we have achieved good results in trying to identify the relation a preposition denotes between two noun phrases, by using their ontological type. However, there is apparently no improvement by including the path from the ontological type to the top of the ontology in the learning process.

Introduction

In traditional search engines information retrieval relies almost entirely on keyword recognition. In the OntoQuery[1] project, we instead match the *conceptual content* of the search phrase and the texts in the database (Andreasen *et al.*, 2002, 2004). In brief, what is done is that concepts are identified through their syntactic form, and mapped into an ontology. The use of ontologies makes it possible to retrieve related concepts to a search phrase, if nothing matches the exact phrase. If one searches for "vehicles", texts that mention specific kinds of vehicles, such as "trucks" or "cars", will also match the query. However, only simple noun phrases are currently being recognized, which is why we are investigating the possibility for expanding the scope of our concept based analysis by including semantic relations between noun phrases, in order to form more complex concepts; initially by trying to recognize relations denoted by prepositions. Our aim is to show that there is an affinity between a semantic relation and the ontological types of the arguments of the relation. In this paper, we present our preliminary results based on machine learning of relations from a Danish corpus compiled from texts from the domain of nutrition.

Semantic relations

Relations exist between entities referred to in discourse, and can be present at different syntactic levels; across sentence boundaries, or within a sentence, a phrase or a word. Also, the relations can have different arity, and can be denoted by different parts of speech.In the approach described here, we will only consider binary relations denoted by prepositions, such as e.g. *trombosis in the heart*, where the preposition *in* denotes a binary locative relation between *trombosis* and *heart*.

[1] http://www.ontoquery.dk

Corpus and annotation

The idea is to perform supervised machine learning, that will take into account the surface form of the preposition and the head of the surrounding noun phrases (NPs), as well as the ontological type of the NP-heads and the relation denoted by the preposition. On this basis, the algorithm should be able to determine relations denoted by prepositions in unseen text. The corpus that we train on, is a small corpus of approximately 18,500 running words compiled of texts from the domain of nutrition, all deriving from "The Danish National Encyclopedia" Gyldendal (2004). The corpus goes through a preprocessing module, that POS-tags, chunks, and marks up the heads of all NPs. The tags used in the annotation of ontological types and relations are:

SIMPLE-tags. The tags used for the ontological type annotation consist of abbreviations of the types in the SIMPLE top ontology. The tag set consists of 151 tags.

Relation-tags. The tags used for the relation annotation derive from a minimal set of relations that have been used in earlier OntoQuery related work. See Jensen & Nilsson (2006); Madsen, Pedersen, & Thomsen (2000, 2001); Nilsson (2001). The final tag set consists of 11 tags.

The preprocessing was done automatically, the ontological type-annotation semi-automatically, whereas the relation annotation was done manually by just one annotator for this initial project. The ideal situation would be to have several annotators annotate the corpus.

Experiments

Our first hypothesis was that there is consistency in which relations prepositions usually denote in particular contexts,

and hence the learning algorithms should be able to generalize well. Our second hypothesis was that the addition of the ontological types of the NP-heads would be the best source of learning. These hypotheses have been tested, and the results were promising (Lassen & Terney, 2006). We were able to achieve a precision of 88.3 using a Support Vector Machine algorithm, SMO (Keerthi *et al.*, 2001), that learned on all information available in the training set. As proposed in Lassen & Terney (2006), we have now expanded the input vector with the full path from the concept corresponding to the NP-head, to the top node in the ontology, in the anticipation that it will improve the precision.

Table 1 below shows the results of the experiments on a data set with 952 instances. The data set it characterized by being both skew and sparse with respect to both relations and ontological types of the corresponding NP-heads. The last column shows the precision of a projected classifier where it outperforms the trivial rejector, which achieves a precision of 37.8%.

Feature space		full path	ontotype	PC
1	Preposition	68.5	68.5	67.6
2	Ontological types	76.3	77.0	61.8
3	Lemma	73.3	73.3	–
4	Lemma and Preposition	83.4	83.4	–
5	Ontological types and Lemma	80.8	81.7	–
6	Ontological types and Preposition	85.5	86.6	–
7	Ontological types, Preposition and Lemma	87.2	88.3	–

Table 1: The precision with SVM when using the full path, the ontological type and a projected classifier on the seven different combinations of input features. "Lemma" here is short for lemmatized NP head.

We find it surprising that the addition of the path from the ontological types of the NP-heads to the top actually degrades performance, though the degration is not statistically significant. However, the richer feature space, which is archived by adding the path, may result in overfitting the data because of the data sparseness and skewness. Please note that line 1,3 and 4 show identical results for both the "full path" and the "ontotype" column, since no ontological knowledge is used in the training.

Conclusion and future work

Even though our experiments are still in an early phase, the results indicate that it is possible to analyse the semantic relations denoted by prepositions using machine learning, an ontology and an annotated corpus – at least within the domain covered by the ontology. Addition of the path to the top in the learning process apparently does not improve the results. Alternative methods of including the path in the learning process could be considered. Future work will include annotation and investigation of a larger specialized corpus, as well as a general language corpus. Also, a more thorough examination of the corpus, more specifically an investigation of which relations or prepositions are most difficult to analyse. Finally, preliminary results suggest that we can actually to some extent predict one ontological type by looking at the other. This apect will also be further investigated.

Acknowledgements

We would like to thank Troels Andreasen, Rasmus Knappe and four anonymous reviewers for fruitful comments.

References

Andreasen, T.; Jensen, P. A.; Nilsson, J. F.; Paggio, P.; Pedersen, B. S.; and Thomsen, H. E. 2002. Ontological extraction of content for text querying. In *Lecture Notes in Computer Science*, volume 2553. Springer-Verlag. 123 – 136.

Andreasen, T.; Jensen, P. A.; Nilsson, J. F.; Paggio, P.; Pedersen, B. S.; and Thomsen, H. E. 2004. Content-based text querying with ontological descriptors. *Data & Knowledge Engineering* 48(2):199–219.

Gyldendal. 2004. The Danish National Encyclopedia. ISBN: 8702031051.

Jensen, P. A., and Nilsson, J. F. 2006. *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*. Springer. chapter Ontology-Based Semantics for Prepositions.

Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; and Murthy, K. R. K. 2001. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation* 13(3):637–649.

Lassen, T., and Terney, T. V. 2006. An onrology-based approach to disambiguation of semantic relations. In *Proceedings of the EACL 2006 workshop on Learning Structured Information in Natural Language Applications (to appear)*.

Madsen, B. N.; Pedersen, B. S.; and Thomsen, H. E. 2000. Semantic relations in content-based querying systems: a research presentation from the ontoquery project. In Simov, K., and Kiryakov, A., eds., *Ontologyes and Lexical Knowledge Bases. Proceedings of the 1st International Workshop, OntoLex 2000*. University of Southern Denmark, Kolding.

Madsen, B. N.; Pedersen, B. S.; and Thomsen, H. E. 2001. Defining semantic relations for ontoquery. In Jensen, P. A., and Skadhauge, P., eds., *Proceedings of the First International OntoQuery Workshop Ontology-based interpretation of NP's*. University of Southern Denmark, Kolding.

Nilsson, J. F. 2001. A logico-algebraic framework for ontologies, ontolog. In Jensen, and Skadhauge., eds., *Proceedings of the First International OntoQuery Workshop Ontology-based interpretation of NP's*. University of Southern Denmark, Kolding.