An Ontology-Based Approach to Disambiguation of Semantic Relations

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

Abstract

This paper describes experiments in using machine learning for relation disambiguation. There have been succesfuld experiments in combining machine learning and ontologies, or light-weight ontologies such as WordNet, for word sense disambiguation. However, what we are trying to do, is to disambiguate complex concepts consisting of two simpler concepts and the relation that holds between them. The motivation behind the approach is to expand existing methods for content based information retrieval. The experiments have been performed using an annotated extract of a corpus, consisting of prepositions surrounded by noun phrases, where the prepositions denote the relation we are trying disambiguate. The results show an unexploited opportunity of including prepositions and the relations they denote, e.g. in content based information retrieval.

1 Introduction

What we describe in this paper, which we refer to as relation disambiguation, is in some sense similar to word sense disambiguation. In traditional word sense disambiguation the objective is to associate a distinguishable sense with a given word (Ide and Véronis, 1998). It is not a novel idea to use machine learning in connection with traditional word sense disambiguation, and as such it is not a novel idea to include some kind of generalization of the concept that a word expresses in the learning task either (Yarowsky, 1992). Other projects have used light-weight ontologies such as WordNet in this kind of learning task (Voorhees, 1993; Agirre and Martinez, 2001). What we believe is our contribution with this work is the fact that we attempt to learn complex concepts that consist of two simpler concepts, and the relation that holds between them. Thus, we start out with the knowledge that some relation holds between two concepts, which we could express as REL(concept1,concept2), and what we aim at being able to do is to fill in a more specific relation type than the generic REL, and get e.g. POF(concept1,concept2) in the case where a preposition expresses a partitive relation. This makes it e.g. possible to determine from the sentence “France is in Europe” that France is a part of Europe. As in word sense disambiguation we here presuppose a finite and minimal set of relations, which is described in greater detail in section 2.

The ability to identify these complex structures in text, can facilitate a more content based information retrieval as opposed to more traditional search engines, where the information retrieval relies more or less exclusively on keyword recognition. In the OntoQuery project[1], pertinent text segments are retrieved based on the conceptual content of the search phrase as well as the text segments (Andreasen et al., 2002; Andreasen et al., 2004). Concepts are here identified through their corresponding surface form (noun phrases), and mapped into the ontology. As a result, we come from a flat structure in a text to a graph structure, which describes the concepts that are referred to in a given text segment, in relation to each other.

However, at the moment the ontology is strictly a subsumption-based hierarchy and, further, only relatively simple noun phrases are recognized and mapped into the ontology. The work presented here expands this scope by including other semantic relations between noun phrases. Our first experiments in this direction have been an analysis of prepositions with surrounding noun phrases (NPs). Our aim is to show that there is an affinity between the ontological types of the NP-heads and the relation that the preposition denotes, which can be used to represent the text as a complex semantic structure, as opposed to simply running text. The approach to showing this has been to annotate a corpus and use standard machine learning methods on this corpus.

2 Semantic relations

The following account is based on the work of (Jensen and Nilsson, 2006): Relations exist between entities referred to in discourse. They can exist at different syntactic levels; across sentence boundaries as in example 1, or within a sentence, a phrase or a word. The relations

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

can be denoted by different parts of speech, such as a verb, a preposition or an adjective, or they can be implicitly present in compounds and genitive constructions as in example 2.

Semantic relations are n-ary: In example 1 below the verb form 'owns' denotes a binary relation between *Peter* and *a dog*, and in example 3, the verb form 'gave' denotes a ternary relation between *Peter*, *the dog* and *a bone*. In example 4 the preposition 'in' denotes a binary relation between *the dog* and *the yard*.

(1) Peter owns a dog. It is a German shepherd.

(2) Peter's dog.

(3) Peter gave the dog a bone.

(4) The dog in the yard.

In the framework of this machine learning project, we will only consider binary relations denoted by prepositions. A preposition, however, can be ambiguous in regard to which relation it denotes. As an example, let us consider the Danish preposition *i* (Eng: in): The surface form *i* in 'A i B' can denote at least five different relations between A and B:

1. A patient relation *PNT*; a relation where one of the arguments' case role is patient, e.g. *"ændringer i stofskiftet"* (changes in the metabolism).
2. A locational relation *LOC*; a relation that denotes the location/position of one of the arguments compared to the other argument, e.g. *"skader i hjertemuskulaturen"* (injuries in the heart muscle).
3. A temporal relation *TMP*; a relation that denotes the placement in time of one of the arguments compared to the other, e.g. *"mikrobiologien i 1800-tallet"* (microbiology in the 19th century).
4. A property ascription relation *CHR*; a relation that denotes a characterization relation between one of the arguments and a property, e.g. *"antioxidanter i renfremstillet form"* (antioxidants in a pure form)
5. A 'with respect to' relation *WRT*; an underspecified relation that denotes an 'aboutness' relation between the arguments, e.g. *"forskelle i saltindtagelsen"* (differences in the salt intake) .

As presented above, the idea is to perform supervised machine learning, that will take into account the surface form of the preposition and the ontological type of the heads of the surrounding noun phrases, and on this basis be able to determine the relation that holds between noun phrases surrounding a preposition in unseen text.

3 The corpus

In order to establish a training set, a small corpus of approximately 18,500 running words has been compiled from texts from the domain of nutrition and afterwards annotated with the ontological type of the head of the noun phrases, and the semantic relation denoted by the preposition [2].

All the text samples in this corpus derive from "The Danish National Encyclopedia" (Gyldendal, 2004), and are thus not only limited domain-wise, but also of a very specific text type which can be classified as expert-to-non-expert. Thus, we cannot be certain that our results can be directly transferred to a larger or more general domain, or to a different text type. This aspect would have to be empirically determined.

3.1 Annotation

For the purpose of learning relations, 952 excerpts of the form:

$$NP - P - NP \tag{5}$$

have been extracted from the corpus and annotated with information about part of speech, ontological type and relation type for NP heads and prepositions, respectively. An example of the analyzed text excerpts are given in table 1 on the following page, where each row indicates a level of the analysis.

The POS-tagging and head extraction have been done automatically, the ontological type assignation partly automatically (ontology look-up) and partly manually (for words that do not exist as instantiations of concepts in the ontology). The relation annotation has been done manually.

The tags used in the annotation on the three levels are:

POS-tags. Our tagger uses a subset of the PAROLE tag set, consisting of 43 tags, see (Hansen, 2000), which means that it is a low level POS tagging with little morphosyntactic information. We only use the tags in order to extract NPs and prepositions, and thus do not need a more fine-grained information level.

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. The set can be seen in table 2

[2] Extraction, POS-tagging and initial ontological and relation type annotation was done by Dorte Haltrup Hansen, CST, University of Copenhagen

surface form	*blodprop* (thrombosis)	*i* (in)	*hjertet* (the heart)
syntactic structure	head of first NP	preposition	head of second NP
relation and ontological type	disease	location	body part

Table 1: Example of the text excerpts analyzed in our experiments. Each row indicate a level of analysis

The manual relation annotation has been done by one annotator for this initial project. The ideal situation would be to have several annotators annotate the corpus. If two or more people annotate the same corpus, they are almost certain to disagree on some occasions. This disagreement can have two sources: first it can be due to cognitive differences. Two people subjected to the same utterance are not guaranteed to perceive the same content, or to perceive the content intended by the producer of the utterance. Many factors are at play here; cultural background, knowledge, memory, etc.

Secondly, it can be due to conceptual, lexical or syntactic ambiguity in the utterance. We cannot remove these sources of disagreement, but we can introduce tools that make the annotation more consistent. By using a finite and minimal realtion tag set and, further, by introducing paraphrase tests, we hope to minimize the risk of inter-annotator disagreement in a future annotation on a larger scale.

3.1.1 The ontological type annotation

As noted above, the ontological types used in the experiments derive from the SIMPLE top ontology (Pedersen, 1999; Lenci et al., 2000). The heads of the phrases have been annotated with the lowest possible node, i.e. ontological type, of the top ontology. In the case of *blodprop* the annotation of ontological type is “disease”, since “disease” is the lowest node in the top ontology in the path from thrombosis to the top. This is illustrated in figure 1, which shows the path from *blodprop* (thrombosis) to the top level of SIMPLE.

Thus, for the purpose of this project, we only consider one node for each concept: the lowest possible node in the top ontology. Another approach would be to consider the the full path to the top node, and also including the path from the leaf node to the lowest node in the top ontology. In the example depicted in figure 1, the full path from trombosis to the top node would be *trombosis–cardiovascular disease–disease–phenomenon–event–entity–top* or *trombosis–cardiovascular disease–disease–agentive–top*.

3.1.2 The set of relations

For the purpose of the manual relation annotation, we needed to decide on a finite set of possible relations that can be denoted by prepositions. This is a non-trivial task, as it is almost impossible to foresee which relations prepositions *can* denote generally, and in the text type at hand specifically, by introspection alone. The method that we decided to use was the following: An initial set of relations that have all been used in prior OntoQuery-related work (Nilsson, 2001; Madsen et al., 2001; Madsen et al., 2000), were chosen as a point of departure. The final set was found by annotating the text segments using this set as the possible relation types, and the relations that are actually manifested in the data then form the final subset that was used as input for a machine learning algorithm. The final subset is shown in table 2.

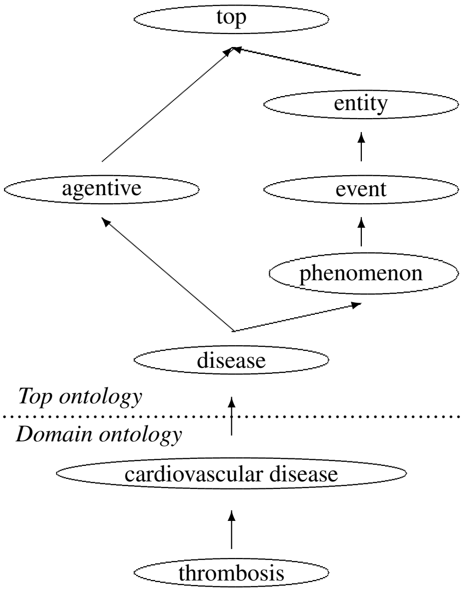

Figure 1: An illustration of the path from *blodprop* (thrombosis) to the top level of the SIMPLE ontology.

Role	Description
AGT	Agent of act or process
BMO	By means of, instrument, via
CBY	Caused by
CHR	Characteristic (property ascription)
CMP	Comprising, has part
DST	Destination of moving process
LOC	Location, position
PNT	Patient of act or process
SRC	Source of act or process
TMP	Temporal aspects
WRT	With respect to

Table 2: The set of relations used in the annotation, which is a subset of the set proposed in Nilsson, 2001.

3.2 Paraphrase tests

In order to ensure a consistent relation annotation, it is necessary to develop a set of paraphrase tests that can help the annotator determine which relation a given preposition denotes in a given context. Some relations are particularly difficult to intuitively keep apart from closely related relations. One of these problematic relation pairs is treated in some detail below.

For example locative and partitive relations can be difficult to keep apart, probably because they to some extent are overlapping semantically. From a philosophical point of view, an important question is 'when does an entity become part of the entity it is located in?', but from a practical point of view, we are interested in answering the question 'how can we decide if a given relation a locative or partitive relation?'.

In this paper we will only treat the latter question. A tool that is useful for this purpose is the paraphrase test: If we can paraphrase the text segment in question into the phrasing the test prescribes, while preserving the semantic content, we can conclude that the relation is a possible relation for the given phrase.

3.2.1 Attribute Transportation Test

The two relations LOC and POF can be difficult to differentiate, even when using paraphrase tests. Therefore, an additional test that could be considered, is Ruus' attribute transportation test (Ruus, 1995)[3]. In the example "The pages in the book", the book gets e.g. the attribute 'binding: {hardback | paperback}' from *cover*, and the attribute 'paper grade:{bond | book | bristol | newsprint}' from *pages*.

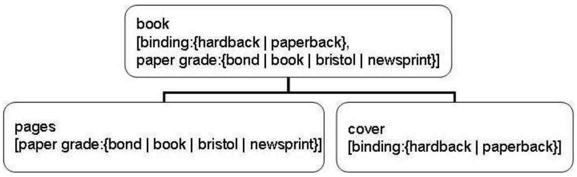
book
[binding:{hardback | paperback},
paper grade:{bond | book | bristol | newsprint}]
pages
[paper grade:{bond | book | bristol | newsprint}]
cover
[binding:{hardback | paperback}]

Figure 2: A graphical representation of the relation between book and pages

We cannot observe an attribute transport, neither from the bird to the roof, nor the other way. This suggests that it is possible to use the atrribute transportation test in order to determine whether a given relation is a POF or a LOC relation. Thus, we can now formulate the following paraphrase test for POF:

POF: *A consists e.g. of B* **and**
A has the attribute X, from B.

4 Experiments

The annotation process generates af a feature space of six dimensions, namely the lemmatized form of the two heads of the noun phrases, the ontological types of the heads, the preposition and the relation. In the corpus there is a total of only 952 text segments. In general the distribution of the data is highly skewed and sparseness is a serious problem. More than half of the instances are of the relation type WRT or PNT, and the rest of the instances are distributed among the remaining 10 relations with only 14 instances scattered over the tree smallest classes. This is illustrated in figure 3. There are 332 different combinations of ontological types where 197 are unique. There are 681 different heads and 403 of them are unique, with all of them being lemmatized.

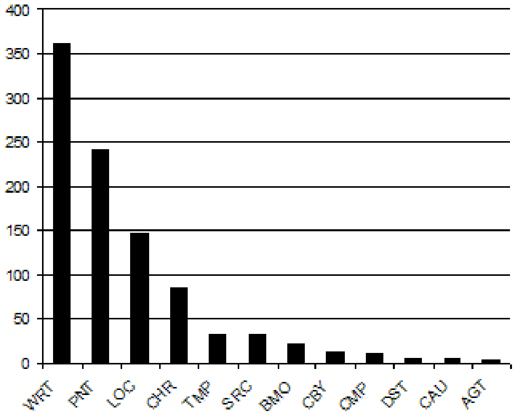

Figure 3: An illustration of the distribution of the 12 possible relations.

Our assumption is that there is consistency in which relations prepositions usually denote in particular contexts, and hence the learning algorithms should be able to generalize well. We also assume that the addition of the ontological types of the head of the NP, is the most vital information in classifying the relation type, at least in this case where data is sparse.

We have run the experiments with a Support Vector Machine algorithm SMO (Keerthi et al., 2001) and the prepositional rule learning algorithm JRip (Cohen, 1995). The former in order to get high precision, the latter in order to get easily interpretable rules for later analysis (see section 4.1). The experiments were run using 10-fold-cross-validation, with a further partition of the training set at each fold into a tuning and a training set. The tuning set was used to optimize the parameter[4] settings for each algorithm . The implementation of the algorithms that we used, was the WEKA software package (Frank et al., 2005).

[3] We will here ignore the question of direction of transport

[4] For SMO the parameters where complexity, kernel used and gamma for the RBF kernel. For JRip it was number of folds used for growing and pruning, minimum number of instances covered and number of optimization runs

The experiments were run on seven different combinations of the feature space, ranging from using only the heads to using both heads, preposition and ontological types of the heads. This was done in order to get insight into the importance of using ontological types in the learning. The results of these experiments are shown in table 3. The last column shows the precision for a projected classifier (PC) in the cases where it outperforms the trivial rejector. The projected classifier, in this case, assigns the relation that is most common for the corresponding input pair; e.g if the ontological types are DIS/HUM, then the most common relation is PNT. The trivial rejector, which assigns the most common relation, in this case WRT, to all the instances, achieves a precision of 37.8%.

Feature space		JRip	SVM	PC
1	Preposition	68.4	68.5	67.6
2	Ontological types	74.4	77.0	61.8
3	Lemma	66.8	73.3	–
4	Lemma and Preposition	72.3	83.4	–
5	Ontological types and Lemma	74,7	81.7	–
6	Ontological types and Preposition	82.6	86.6	–
7	Ontological types, Preposition and Lemma	84,0	88.3	–

Table 3: The precision of SVM, JRip and a projected classifier on the seven different combinations of input features. "Lemma" here is short for lemmatized NP head.

The following conclusions can be drawn from table 3. The support vector machine algorithm produces a result which in all cases is better than the baseline, i.e. we are able to produce a model that generalizes well over the training instances compared to the projected classifier or the trivial rejector. This difference is not statistically significant at a confidence level of 0.95 when only training on the surface form of prepositions.

A comparison of line 1–3 shows that training on ontological types seems to be superior to using lemmatized NP heads or prepositions, though the superiority is not statistically significant when comparing to the lemmatized NP heads. When comparing line 4–7 the difference between the results are not statistically significant. This fact may owe to the data sparseness. However, comparing line 1 to line 6 or 7, shows that the improvement of adding the preposition and the lemmatized NP heads to the ontological types is statistically significant.

In general, the results reveal an unexplored opportunity to include ontological types and the relations that prepositions denote in information retrieval. In the next section, we will look more into the rules created by the JRip algorithm from a linguistic point of view.

4.1 Analyzing the rules

In this section we will take a deeper look into the rules produced by JRip on the data set with only ontological types, since they are the most interesting in this context.

The JRip algorithm produced on average 21 rules. The most general rule covering almost half of the instances is the default rule, that assigns all instances to the WRT relation if no other rules apply. At the other end of the spectrum, there are ten rules covering no more than 34 instances, but with a precision of 100%. It is futile to analyse these rules, since they cover the most infrequent relations and hence may be overfitting the data set. However, this seems not be the case with a rule like "*if* the ontotype of the first head is DISEASE and and the ontotype of the second head is HUMAN *then* the relation is PATIENT" covering an instance as e.g. "iron deficiency in females".

The rule with the second highest coverage, and a fairly low precision of around 66%, is the rule: "*if* the ontotype of the second head is BODY PART *then* the relation type is LOCATIVE". The rule covers instances as e.g. "...thrombosis in the heart" but also incorrectly classifies all instances as LOCATIVE where the relation type should be SOURCE. E.g. the sentence '...iron absorbtion from the intestine", which is in fact a SOURCE relation, but is classified as LOCATIVE by the rule.

One of the least surprising and most precise rules is: "*if* the ontotype of the second head is TIME *then* the relation type is TEMPORAL" covering an instance as e.g. "...diet for many months". We would expect a similar rule to be produced, if we had performed the learning task on a general language corpus.

5 Conclusion and future work

Even though the experiments are in an early phase, the results indicate that it is possible to analyse the semantic relation a preposition denotes between two noun phrases, by using machine learning and an annotated corpus – at least within the domain covered by the ontology. Future work will therefore include annotation and investigation of a general language corpus. Also, a more thorough examination of the corpus, more specifically an investigation of which relations or prepositions that are most difficult to analyse. Also, we will experiment with the amount of information that we train on, not as we have already done by in- or excluding *types* of information, but rather the extension of the information: Could we predict the ontological type of one of the arguments by looking at the other? Finally, an explicit inclusion of the whole ontology in the learning process is on the agenda, as proposed in section 3.1.1 on page 3, in the anticipation that the learner will produce an even better model.

6 Acknowledgements

We would like to thank Troels Andreasen, Per Anker Jensen and two anonymous reviewers for fruitful comments. The latter especially for comments on the experimental part and inter-annotator agreement.

References

[Agirre and Martinez2001] E. Agirre and D. Martinez. 2001. Learning class-to-class selectional preferences.

[Andreasen et al.2002] Troels Andreasen, Per Anker Jensen, Jørgen Fischer Nilsson, Patrizia Paggio, Bolette Sandford Pedersen, and Hanne Erdman Thomsen. 2002. Ontological extraction of content for text querying. In *Lecture Notes in Computer Science*, volume 2553, pages 123 – 136. Springer-Verlag.

[Andreasen et al.2004] Troels Andreasen, Per Anker Jensen, J&#248;rgen Fischer Nilsson, Patrizia Paggio, Bolette Sandford Pedersen, and Hanne Erdman Thomsen. 2004. Content-based text querying with ontological descriptors. *Data & Knowledge Engineering*, 48(2):199–219.

[Cohen1995] William W. Cohen. 1995. Fast effective rule induction. In Armand Prieditis and Stuart Russell, editors, *Proceedings of the 12th International Conference on Machine Learning*, pages 115–123, Tahoe City, CA. Morgan Kaufmann.

[Frank et al.2005] Eibe Frank, Mark Hall, and Len Trigg. 2005. Weka. Publicly available, November.

[Gyldendal2004] Gyldendal. 2004. The danish national encyclopedia. ISBN: 8702031051.

[Hansen2000] Dorte Haltrup Hansen. 2000. Træning og brug af brill-taggeren på danske tekster. Technical report, CST.

[Ide and Véronis1998] Nancy Ide and Jean Véronis. 1998. Special issue on word sense disambiguation: Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24.

[Jensen and Nilsson2006] Per Anker Jensen and Jørgen Fischer Nilsson, 2006. *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*, chapter Ontology-Based Semantics for Prepositions. Springer.

[Keerthi et al.2001] S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649.

[Lenci et al.2000] Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari1, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

[Madsen et al.2000] Bodil Nistrup Madsen, Bolette Sandford Pedersen, and Hanne Erdman Thomsen. 2000. Semantic relations in content-based querying systems: a research presentation from the ontoquery project. In K Simov and A Kiryakov, editors, *Ontologyes and Lexical Knowledge Bases. Proceedings of the 1st International Workshop, OntoLex 2000*. University of Southern Denmark, Kolding.

[Madsen et al.2001] Bodil Nistrup Madsen, Bolette Sandford Pedersen, and Hanne Erdman Thomsen. 2001. Defining semantic relations for ontoquery. In Per Anker Jensen and P Skadhauge, editors, *Proceedings of the First International OntoQuery Workshop Ontology-based interpretation of NP's*. University of Southern Denmark, Kolding.

[Nilsson2001] Jørgen Fischer Nilsson. 2001. A logico-algebraic framework for ontologies, ontolog. In Jensen and Skadhauge, editors, *Proceedings of the First International OntoQuery Workshop Ontology-based interpretation of NP's*. University of Southern Denmark, Kolding.

[Pedersen1999] Bolette Sandford Pedersen. 1999. Den danske simple-ordbog. en semantisk, ontologibaseret ordbog. In C. Poulsen, editor, *DALF 99, Datalingvistisk Forenings årsmøde 1999*. Center for sprogteknologi.

[Ruus1995] Hanne Ruus. 1995. *Danske kerneord. Centrale dele af den danske leksikalske norm 1-2*. Museum Tusculanums Forlag.

[Voorhees1993] Ellen M. Voorhees. 1993. Using wordnet to disambiguate word senses for text retrieval. In Robert Korfhage, Edie M. Rasmussen, and Peter Willett, editors, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, June 27 - July 1, 1993*, pages 171–180. ACM.

[Yarowsky1992] David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes, France, July.