An Ontology-Based View on Prepositional Senses

Tine Lassen
Roskilde University
Denmark
tlassen@ruc.dk

Abstract

This paper describes ongoing work, aimed at producing a lexicon of prepositions, i.e. relations denoted by prepositions, to be used for information retrieval purposes. The work is ontology based, which for this project means that the ontological types of the arguments of the preposition are considered, rather than the word forms. Thus, sense distinctions are made based on ontological constraints on the arguments.

1 Introduction

In traditional web search engines, information retrieval relies more or less exclusively on simple string match. In the OntoQuery project[1], ontology-based search in text databases is performed based on a match between the *conceptual content* of the search phrase and the text segments in the database.(Andreasen et al., 2002; Andreasen et al., 2004). In short, concepts are identified through their corresponding surface form and mapped into an ontology. The use of an ontology makes it possible to introduce the notion of conceptual distance and thereby ranking the search result by semantic similarity. E.g. “pony” and “zebra” may be more similar concepts than “pony” and “lion”, because the distance when traversing a graph representation of the ontology is longer going from “pony” to “lion” than from “pony” to “zebra”. See figure 1 for a simplified excerpt of the ontology.

However, only relatively simple noun phrases are currently recognized and mapped into the ontology, and we are thus investigating the possibilities of expanding the scope of our concept-based analysis by including the semantic relations that hold between noun phrases. Our first experiments in this have been an analysis of prepositions and their surrounding noun phrases. The immediate aim is to be able to include a lexicon of prepositions, consisting of a lexicon entry for each sense of a given preposition (a sense, in this context, is a relation that it can denote). Each entry has an argument structure and ontological constraints on the arguments. Thus, a preposition in a given context will be assigned a pertinent sense based on the ontological types of its surrounding noun phrases. If this can be achieved, we will be able to say, e.g., that the text segments “she is riding *her pony in the morning*” and “he was riding *a pony during the war*” are more closely related, i.e. the relational distance is smaller, than any of the two to “she was riding *her pony in the hall*”. The relations that hold between “pony” and “morning”, and “pony” and “war”, denoted by *in* and *during* respectively, are of a temporal nature, whereas the relation that holds between “pony” and “hall”, denoted by *in*, is of a locative nature. The notion of relational distance is similar to that of conceptual distance; the distance when traversing a graph representation of the relation ontology. A combined measure, that takes into account both the conceptual and rela-

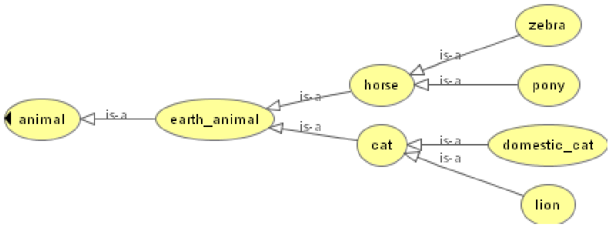

Figure 1: Excerpt from the ontology with the concepts ’horse’ and ’cat’

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

Proceedings of the Third ACL-SIGSEM Workshop on Prepositions, pages 45–50,
Trento, Italy, April 2006. ©2006 Association for Computational Linguistics

tional distance would have to be introduced in order to express the combined distance between such more complex structures (structures that are more complex than the simple noun phrases, that is), but this question is beyond the scope of this paper[2].

Initially, we use a predefined set of relations that was originally proposed in (Nilsson, 2001) (see table 1). This set is hierarchically unstructured, which means that per default, the conceptual distance between any given two pairs of relations is the same. We will later introduce an ontology of relations, which will make it possible to express that the distance between e.g. the partitive and the locative relation is smaller than the distance between the locative and the causal relation (see section 4.1).

2 Approach

We are using a bottom-up approach, in which we manually annotate a corpus[3] with semantic relations for all occurences of prepositions that are surrounded by noun phrases. Further we annotate the heads of the surrounding noun phrases with their ontological type and subsequently analyze the result in order to produce ontological constraint rules. The ontology that was used for the ontological type annotation, is the SIMPLE top ontology (Pedersen, 1999; Lenci et al., 2000).

Relations exist between the entities referred to in discourse, and can exist at different syntactic levels; across sentence boundaries as in *Peter owns a pony. It is stubborn* , or whithin a sentence, a phrase or a word. The relations 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 *Peter's pony*.

The following account is based on the work of (Jensen and Nilsson, 2006): Semantic relations are n-ary (where n$\geq$1): In the example *Peter owns a pony* the verb 'owns' denotes a binary relation between *Peter* and *a pony*, and in example *Peter gave the pony a carrot*, the verb 'give' denotes a ternary relation between *Peter*, *the pony* and *a carrot*. In the example *The pony in the field* the preposition 'in' denotes a binary relation between *the pony* and *the field*. In the framework of this project, however, we will only consider binary relations denoted by prepositions. Using the algebraic description language OntoLog (Nilsson, 2001), we express binary relations as *A[REL:B]*, where the first argument of the relation, *A*, relates to the second argument, *B*, in the manner *REL*.

A preposition, however, can be ambiguous in regard to which relation it denotes (we assume a restricted set of possible relations for prepositions, see table 1). 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. *"generalforsamlingen i 1981"* (the general assembly in 1981).

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) .

Role	Description
TMP	Temporal aspects
LOC	Location, position
PRP	Purpose, function
WRT	With respect to
CHR	Characteristic (property ascription)
CUM	Cum (i.e., with, accompanying)
CBY	Caused by
CAU	Causes
BMO	By means of, instrument, via
CMP	Comprising, has part
POF	Part of
AGT	Agent of act or process
PNT	Patient of act or process
SRC	Source of act or process
RST	Result of act or process
DST	Destination of moving process

Table 1: The set of possible relations used in the annotation process (Nilsson, 2001)

[2] For a discussion of a distance measure between noun phrases, see e.g. (Bulskov and Andreasen, 2004) or (Knappe and Andreasen, 2002)

[3] The corpus is a small corpus of approximately 18,500 running words has been compiled from texts from the domain of nutrition.

3 Results

Following the initial annotation, we performed an analysis of all occurences of the relations and the ontological types of their arguments. Could we identify patterns that could result in lexical rules for the lexicon? The limited space here does not allow us to show the full results of the analysis, so we will focus on one preposition, the Danish preposition *i* (Eng: *in*) and later focus on one relation type denoted by that preposition, namely the locative relation. There are 199 occurences of the preposition *i* in the corpus, and the relations that it denotes are distributed as follows:

LOC (137/199 : 68,8%)
WRT (25/199 : 12,5%)
TMP (17/199 : 8,5%)
PNT (11/199 : 5.5%)
CHR (9/199 : 4,5%)

If we look at the LOC relation, which is the most frequent relation denoted by *i* in the corpus, we get this distribution of ontological types for the arguments: Of the 137 instances of *i* denoting a locative relation, there are 57 different ontological type-pairs, if we consider unique occurences of a given onlogical type-pair (a pair, meaning the ontological types of the two arguments combined), 31 different first arguments, and 16 different second arguments. The most significant ontological type for arguments is the type "body part" (BPA), which occurs 10 times as first argument, and 66 times as second argument. However, in total, the type occurs 119 times (13 times as first argument and 106 times as second argument) in the corpus as a whole. If we were to implement a rule that would assign the relation LOC to any preposition that has the ontological type "body part" as any argument, we would get a precision[4] score of 68.9., a higher score of 92.3 if we only consider the first argument, and 66 if we only consider the second argument.

However, if we limit the rule to assign the relation LOC only to occcurences of the preposition *i*, with arguments of the type BPA, then we get a precision score of 100. This sounds promising, but it should be noted that the coverage of the best rule "*IF* any argument is BPA *AND* preposition is 'i' *THEN* assign LOC to preposition" is quite low: the recall[5] score for the rule is 55.8, which means that we can correctly assign a relation to 55.8% of the LOC senses of the preposition *i*, and these only make up 68.8% of the total number of *i*-occurences. In fact, we can only correctly assign the LOC relation to 38.7% of the actual relations denoted by *i* in the corpus using this rule. Only if we can produce more rules of this type with high precision scores, we can be optimistic about the outcome of the project.

Rule	Precision	Recall

IF any argument is BPA
THEN assign LOC to preposition
	68.9	56.5

IF first argument is BPA
THEN assign LOC to preposition
	92.3	8.3

IF second argument is BPA
THEN assign LOC to preposition
	66	48.3

IF any argument is BPA
AND preposition is 'i'
THEN assign LOC to preposition
	100	55.8

IF first argument is BPA
AND preposition is 'i'
THEN assign LOC to preposition
	100	7.2

IF second argument is BPA
AND preposition is 'i'
THEN assign LOC to preposition
	100	48.6

Table 2: Precision and recall scores for rules that assign the LOC relation to prepositions with a BPA constraint on the ontological type of arguments

4 Suggestions to improve the results

In the following, we propose a way of improving the results by introducing a relation ontology, and further, by either generalizing or specializing the ontological type level for the arguments. Our hypothesis is that by doing this, we will end up with rules that have broader coverage. By coverage, we mean the number of occurences that the rule applies to, compared to the number of occurences that potentially could be covered by the rule.If the relations are too general, then we miss out on some of the semantic content of the relation between the items that we consider, and we want to capture as much semantics as we can. On the other hand, in some cases it may be that we have made distinctions in the relation set that are not detectable when analyzing the data. Also, if the ontological type distribution for the

[4] Precision = number of correct matches / number of matches

[5] Recall = number of correct matches / ideal number of matches

arguments is too coarse or too fine grained, the patterns that appear when we analyze the data, will not be general enough to produce rules from.

4.1 The relation ontology

The flat list of possible relations, as can be seen in table 1, that we initially used, now has to be transformed into a relation ontology. Our heuristics for doing this, in short, is to group relations that are more closely related than others in subbranches, such that the distance between them is shorter than the distance to other less closely related relations. In figure 2, the intralocal and extralocal relations are more closely related than e.g. the intralocal and dynamic relation, because the distance when traversing the graph is two archs for the former, and three arcs for the latter. One way of deciding relatedness, is to say that if two relations have proven difficult to differentiate in the initial annotation process, then they are probably more closely related. This is the approach we have chosen. Also, we have grouped together other relations, such as the bidirective relations ’part of’ and ’has parts’, and ’causes’ and ’caused by’, and other relations that naturally group together, such as the theta roles ’agent’-’patient’ and ’source’-’result’.

A possible next step is to specialize the relations that *can* be specialized. The relations that intuitively make sense to specialize are the temporal (as has been done in OWL-Time ontology (formerly DAML-Time) (Hobbs and Pan, 2004)), partitive (Winston and Hermann, 1987) and spatial/local relations (e.g. DOLCE Spatialrel ontology)[6].

Our work with specialisations of the relation ontology, particularly the local relations, is largely inspired by Pustejovsky’s work on event structures (Pustejovsky, 1996). Pustejovsky suggests a subdivision of complex events into subevents.

However, another way of expressing the difference between events with one or more subevents is, as we will do in the following, static and dynamic relations: static relations only consist of one subevent, and dynamic relations have more than one subevent.

- A dynamic locative relation is a complex event, that consists of more than one subevent, and it denotes a source or a goal of a process, or a place where a process unfolds.

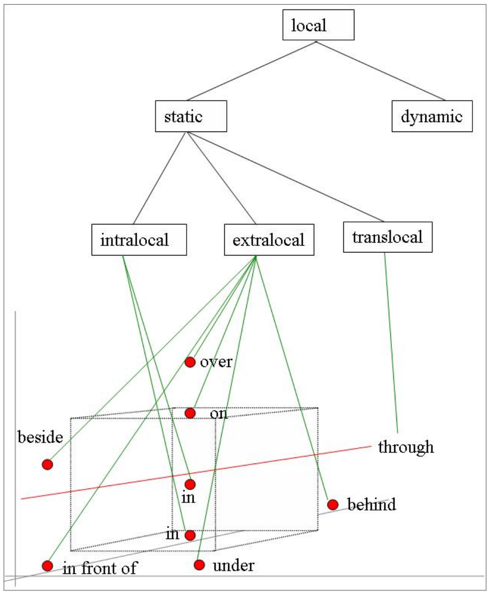

Figure 2: Illustration of static locative relations

- A static locative relation consists of just one subevent, and it denotes ’being located at’.

Another type of specialisation of the locative relation could be a subdivision into relations concerning area, region, distance, etc. (as it has been done in DOLCE), but the aforementioned static and dynamic locative relations appears to be more appropriate when the subject is relations denoted by prepositions.

As a possible further specialisation of dynamic and static relations we suggest:

- Intralocal: an intralocal relation denotes a goal of a process (e.g. *into the box*), or a location within a delimited area.

- Extralocal: an extralocal relation denotes a point of departure of a process (e.g. *out of the box*), or a location outside or touching the outer limitation of a delimited area.

- Translocal: a translocal relation denotes a location or a process through a delimited area (e.g. *through the box*).

[6] http://www.loa-cnr.it/Files/DLPOnts/SpatialRel_397.owl

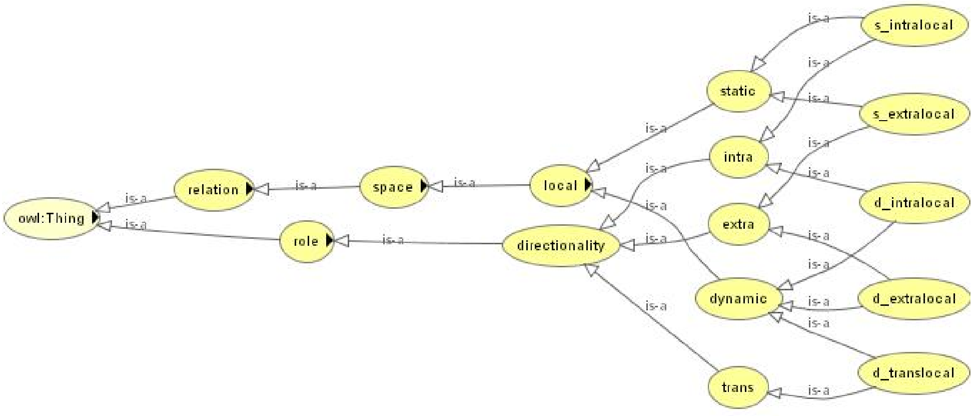

Figure 4: Excerpt of the relation ontology containing the locative relations

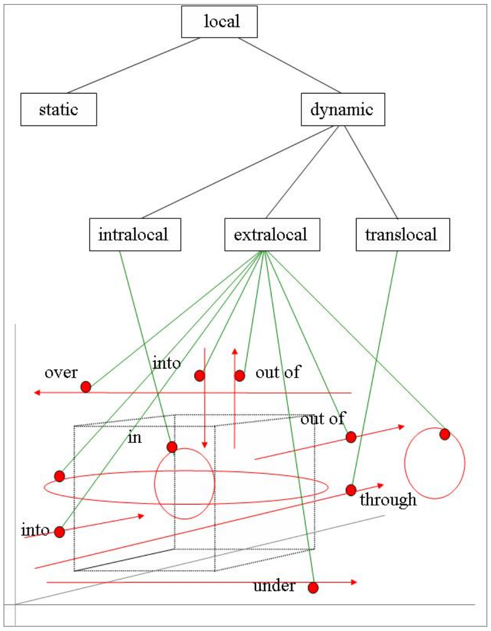

Figure 3: Illustration of dynamic locative relations

We now reannotate the locative relations according to these new subtypes of the locative relation; but conforming to the bottom-up approach, we initially only subdivide into static and dynamic locative relations. Can we observe a clearer pattern with respect to the ontological types of the arguments, if we do this? If so, then we try a more fine grained subdivision.

The results of the LOC dynamic and LOC static subdivision show that of the original 137 instances of *i* that denote a locative relation, 33 denote a LOC dynamic relation, and 104 denote a LOC static relation. The patterns that we observe, are actually clearer: for all, but one, of the dynamic relations the first arguments denote some kind of process or event, whereas the second arguments are all more or less specialized types of concrete entity. The most prevalent ontological type for the static local relation is “natural substance”, which occurs 50 times. For the dynamic local relation, the most prevalent relation for the first argument is “change”, which occurs 17 times. The most prevalent second argument is again “body part”, which also occurs 17 times. If we now calculate precision and recall for rules that constrain arguments of the static and dynamic locative relations to the most prevalent ones, we get the results shown in tables 3 and 4. However, we only calculate scores for rules constraining the first argument, and only for the preposition *i*, because only part of the corpus has been annotated with these relations.

The precision score for the best rule is lower than for the original LOC-rules (92.1 compared to 100). However, considering that we capture more semantics, and the fact that the arguments of the static and dynamic locative relations are more

Rule	Precision	Recall

IF any argument is BPA
AND preposition is ’i’
THEN assign LOC>static to preposition
	64.1	48.1

IF first argument is NSU
AND preposition is ’i’
THEN assign LOC>static to preposition
	92.1	33.7

IF second argument is BPA
AND preposition is ’i’
THEN assign LOC>static to preposition
	68.7	47.1

Table 3: Precision and recall scores for the rule that assign the LOC>static relation to the preposition *i* with constraints on the ontological types the arguments

Rule	Precision	Recall

IF first argument is CHA
AND preposition is ’i’
THEN assign LOC>dynamic to preposition
	58.6	51.5

IF second argument is BPA
AND preposition is ’i’
THEN assign LOC>dynamic to preposition
	25.4	51.5

Table 4: Precision and recall scores for the rule that assign the LOC>dynamic relation to the preposition *i* with constraints on the ontological type of the first and second argument

uniform in their distribution, this indicates that a generalisation of the ontological types of the arguments will result in even better rules, i.e. rules with a larger coverage.

5 Conclusion and further work

Our aim is to show that ontological types can be used as constraints in a lexicon of semantic relations denoted by prepositions. In this paper we have presented our preliminary results, that are based on an analysis of a Danish corpus, compiled of texts from the domain of nutrition. We have introduced an ontology of relations, which will make it possible to measure relational distance between complex structures in addition to the conceptual distance that we can measure between concepts. The results are promising: We can produce rules that have good precision scores for the locative relation, and we expect to improve the rules by generalizing the ontological types of the prepositional arguments. Also, we plan to expand our research to cover other relations than the ones treated in this paper.

6 Acknowledgements

We would like to thank Troels Andreasen, Per Anker Jensen and three anonymous reviewers for fruitful comments and suggestions.

References

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.

Troels Andreasen, Per Anker Jensen, Jørgen 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.

R. Knappe H. Bulskov and T. Andreasen. 2004. Perspectives on ontology-based querying. *International Journal of Intelligent Systems, to appear*.

Jerry R. Hobbs and Feng Pan. 2004. An ontology of time for the semantic web. *ACM Trans. Asian Lang. Inf. Process.*, 3(1):66–85.

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.

H. Bulskov R. Knappe and T. Andreasen, 2002. *Flexible Query Answering Systems*, chapter On Measuring Similarity for Conceptual Querying, pages pp. 100–111. Number 2522 in Lecture Notes in Artificial Intelligence.

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.

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.

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.

James Pustejovsky. 1996. *The generative lexicon*. MIT Press, Cambridge, Mass.

Roger Winston, Morton E. Chaffin and Douglas Hermann. 1987. A taxonomy of part-whole relations. *Cognitive Science*, 11:417–444.