

Semantic Relations between Concepts in Danish Domain Specific Texts

Lone Bo Sisseck

Dept. of Computational Linguistics
Copenhagen Business School
Bernhard Bangs Allé 17 B
DK-2000 Frederiksberg
Tel: +45 38 15 33 61
Fax: +45 38 15 38 20
lbs.id@cbs.dk

Abstract

This paper describes a fragment of an ongoing Ph.D. project that aims at identifying and extracting conceptual relations from Danish domain specific text. From a Danish corpus on nutrition, linguistic patterns indicating the generic relation have been extracted manually on the basis of an arbitrarily selected test corpus. These patterns, ranging over a hybrid of linguistic expressions, have then been tested on the whole corpus and the resulting data reveals an interesting point of departure for further studies. Out of 124 sentences where the linguistic marker actually is a candidate marker for a semantic relation between concepts, i.e. domain specific concepts are actually present in the sentence construction, 93 occurrences indicate a semantic relation between two or more concepts. The concepts have been mapped into a concept system and the result is compared with existing fragments of a manually build ontology.

1. Background

My project is related to the OntoQuery project¹, OQ, which aims at developing ontology based search methods. Since ontology based search systems require huge amounts of domain specific ontologies in order to be a success and since it is a very time consuming task to build an ontology manually it will be of great interest to find ways of automating at least part of the ontology generation. Earlier and ongoing projects throughout the world have already developed various methods of more or less automatic term and relation extraction from both a statistical and a linguistic approach. However, there have hardly been any attempts to investigate Danish texts in order to find linguistic patterns that could indicate relations between concepts.

2. The Generic Relation

In the (manually build) OQ ontology, which is based on the above-mentioned corpus on nutrition, only the generic *is-a* relation has been implemented in the prototype ontology so far. The nutrition ontology has been integrated with the SIMPLE-ontology². The *is-a* relation is also known as the *generic relation* or the *type* relation.

The following definition of the generic relation is taken from the ISO standard on terminology work (ISO 1087-1): relation between two concepts where the intension of one of the concepts includes that of the other concept and at least one additional delimiting characteristic.

The relation between concepts described in the following will be illustrated by the expression $aR(b, \dots, n)$ where a concept, a , is related to one or more concepts, b, \dots, n , by the relation R .

3. The Pilot Project

As a starting point I chose to initiate my investigation by solely looking for the generic relation as in OQ. I selected a few articles from the nutrition corpus, one about nutrition in general and one about Vitamin A. In these articles I located sets of two or more concepts that were related so that one of the concepts was a superordinate concept of the other(s). The linguistic patterns that indicated the relation are outlined out in Table 1.

The whole nutrition corpus has been investigated with the corpus tool WordSmith in order to find out if it is possible to locate semantic relations throughout the whole corpus by using the manually found linguistic markers as search patterns. Since the nutrition corpus is very small, around 20000 words, it was possible to go through every occurrence manually in order to decide whether we are dealing with an indicator of a semantic relation. The following conditions were set up:

- Primarily two or more domain specific concepts should be involved in the collocation pattern
- Secondly I should be able to recognize the presence of a semantic relation
- The relation should be the generic relation

It was then possible to eliminate many of the occurrences as semantic relations due to the fact that no domain specific concepts were involved. The remaining occurrences, i.e. the possible semantic relations, were then manually analyzed in order to see how many of them actually indicated the generic relation between two or more concepts. The results from this analysis are summarized in Table 2.

¹ The purpose of the interdisciplinary project OntoQuery is to develop theories and methods for content-based information retrieval. Partners in the project are the following Danish research institutions: Copenhagen Business School, the Danish Technical University, Roskilde University and the University of Copenhagen. The project is funded by the Danish Research Agency under the Information Technology Programme. For further information please see www.ontoquery.dk

² The Danish SIMPLE lexicon has been developed at the Centre for Language Technology in Denmark.

expression	aR(b,...,n) where R =	collocation example
at være (to be)	er (is)	Chrom er et sporstof (chromium is a tracer)
colon :	:	...vitaminer: B1, ...,n (vitamins: B1,...,n)
parenthesis ()	()	...organisk jern (hæmoglobin, myoglobin) ...(organic iron (haemoglobin, myoglobin))
fx (e.g)	fx (e.g.)	...spormetaller, fx selen og kobber (...trace metals e.g. selenium and cobber)
omfatte (to include)	omfatter (include(s))	...næringsstoffer omfatter vitaminer og mineraler (...nutrients include vitamins and minerals)

Table 1: The *is-a* linguistic markers

expression	aR(b,...,n) where R =	oc- curen- ces #	pos- sible rela- tions #	act- tual gen - eric rela- tions #	act- tual gen - eric rela- tions %
at være (to be)	er (is)	447	65	45	70 %
colon :	:	70	37	29	78 %
parenthesis ()	()	11	9	8	88 %
fx (e.g)	fx (e.g.)	10	9	9	100 %
omfatte (to include)	omfatter (include(s))	4	4	2	50 %
Total			124	93	75 %

Table 2: Evaluation scheme

Note that apart from the 11 occurrences expressed by the parenthesis marker there were 106 occurrences with only one term inside the parenthesis. 84 of these related domain specific concepts e.g. *xerofthalmi (øjentørsot)(xerophthalmia (dry eyes))*. Following the above-mentioned conditions they indicate possible generic relations. However, 76 of the 84 (90%) are synonyms like in the example. The remaining 10% were not synonyms nor were they related concepts. So the 11 occurrences in the evaluation scheme all involve two or more concepts inside the parenthesis.

Out of 124 sentences where the linguistic marker actually is a candidate marker for a semantic relation between concepts, meaning that some domain specific concepts are actually present in the sentence construction, 93

occurrences indicate a semantic relation between two or more concepts, that is 75 %. One could argue that this is not an impressive result and it will be necessary to aim at an accuracy rate much higher. I must emphasize that this is an initial pilot project, and I feel highly motivated by the accuracy rate of 75 % to continue the investigation.

The next step will most certainly be to experiment by applying a statistical method to my training data in order to measure the effect. For that special purpose the naive Bayes classifier approach to Word Sense Disambiguation³ will be considered as a working method. I have chosen this method because the linguistic markers, both signs and words, have different “senses” throughout the text. As I pointed out previously, the parenthesis can have the “generic” sense, the “synonym” sense and maybe also other senses. It would therefore be interesting to improve on the result by applying a statistical method. From my data I wish to estimate the probability of the *is-a* relational sense of each of the linguistic markers in combination with the probabilities of the presence of concepts in the surroundings.

4. The Ontology

The related concepts that have been extracted on the basis of the five linguistic markers can be seen in Table 3. It should be noted that the linguistic markers don’t carry any proven information about the direction of the relation, and there has been no effort made so far to solve this super/subordinate concept problem.

A project at Syddansk Universitet (Weilgaard, 2000) has investigated the suitability of a number of Danish verbs as knowledge probes for the retrieval of definitions in a Danish corpus on Hydraulics. One of the advantages of this method is that when using a definitorial verb as search pattern, meaning a verb that carries some kind of concept definition, a picture emerges of the relational patterns between the involved concepts. That means that the linguistic marker also indicates which concept is superordinate to the other and vice versa.

Despite the super/subordinate concept problem, the concepts have been manually mapped into a concept system. The decision about which concept should be the superordinate concept was based on intuition. In the collocation example from table 1, *chromium is a tracer*, the concept *tracer* is the superordinate concept because intuitively there can be different kinds of *tracers*. In the collocation example, *vitamins (B1,...,n)*, the concept *vitamins* is the superordinate concept because intuitively there are several kinds of *vitamins*.

In order to see if there is any reason in trying to relate concepts solely on the basis of the retrieved linguistic markers, the ontology is compared to the similar fragments from the manually build ontology in the OntoQuery project.

The Ontology built on the basis of the concept relationships Extracted from the Linguistic Markers, referred to as OELM in the following, can be seen in Figure 1.

³ The naive Bayes classifier approach to WSD is based on the premise that choosing the best sense for an input vector amounts to choosing the most probable sense given that vector (Jurafsky & Martin 2000, p. 638).

ernæringsproblem, <i>nutrition problem</i>	sult, <i>hunger</i> underernæring, <i>malnutrition</i>
ernæringsproblem, <i>nutrition problem</i>	overvægt, <i>overweight</i> undervægt, <i>underweight</i> psykiske spiseforstyrrelser, <i>psyk. eating disorder</i>
ernæringsproblem, <i>nutrition problem</i>	A-vitaminmangel, <i>vitamin A deficiency</i>
næringsstof, <i>nutrition substance</i>	zink, <i>zinc</i>
næringsstof, <i>nutrition substance</i>	riboflavin, <i>riboflavin</i> selen, <i>selenium</i> zink, <i>zinc</i> kobber, <i>copper</i> mangan, <i>manganese</i>
næringsstof, <i>nutrition substance</i>	fedt, <i>fat</i> kulhydrat, <i>carbohydrate</i> protein, <i>protein</i>
næringsstof, <i>nutrition substance</i>	vitamin, <i>vitamin</i> mineral, <i>mineral</i>
næringsstof, <i>nutrition substance</i>	kulhydrat, <i>carbohydrate</i> fedt, <i>fat</i> protein, <i>protein</i>
næringsstof, essentiel, <i>essential nutrition substance</i>	natrium, <i>sodium</i> klor, <i>chlorine</i>
næringsstof, essentiel, <i>essential nutrition substance</i>	vitamin, <i>vitamin</i> mineral, <i>mineral</i> sporstoffer, <i>tracer</i>
signalstof	nitrogenoxid, <i>nitrogen</i>
signalstof	serotonin, <i>serotonin</i>
signalstof	adrenalin, serotonin, kreatin
sporstof, <i>tracer</i>	chrom, <i>chromium</i>
sporstof, <i>tracer</i>	jern, <i>iron</i> zink, <i>zinc</i>
sporstof, essentiel, <i>essential tracer</i>	kobber, <i>copper</i> molybdæn, <i>molybdenum</i>
stof, <i>substance</i>	steroid, <i>steroid</i>
stof, <i>substance</i>	selen, <i>selenium</i>
stof, <i>substance</i>	bilirubin, <i>bilirubin</i> ubiguinon, <i>ubiquinon</i> glutathione, <i>glutathion</i>
stof, livsvigtigt, <i>vital substance</i>	magnesium
vitamin, <i>vitamin</i>	A, B1, B2, B6, B12, C, D, E,
vitamin, fedtopløselig, <i>fat soluble vitamin</i>	A, D, E, K

Table 3: related concepts

When I extracted a similar fragment of the manually built OntoQuery ontology, an ontology emerged that showed a certain degree of similarity but it also contains substantial differences. The OntoQuery Ontology Fragment, referred to as OOF in the following, can be seen in Figure 2.

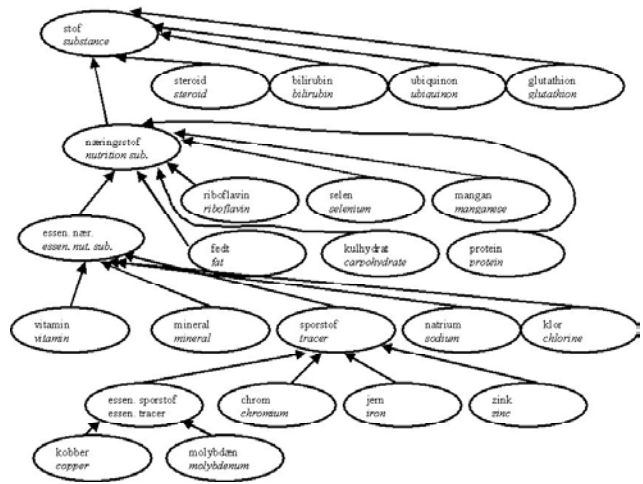


Figure 1: Ontology Extracted from Linguistic Markers, OELM

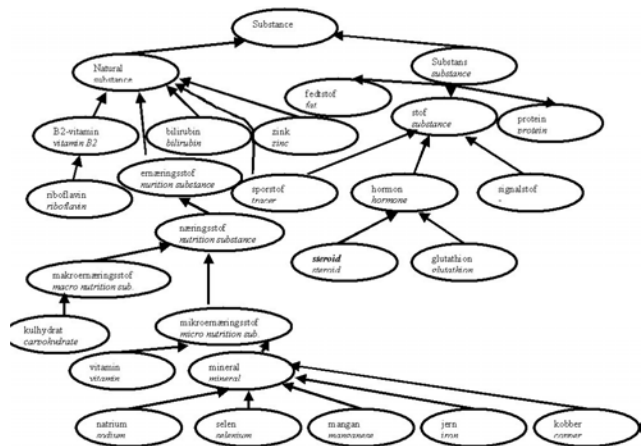


Figure 2: The OntoQuery Ontology Fragment, OOF.

For example, in the OOF there is a “hormon”-level in between the superordinate concept *substance* and the two subordinate concepts *steroid* and *glutathion*, whereas the OELM tends to be more general, e.g. *steroid* and *glutathion* are directly subordinate concepts to *substance* on the same level as *bilirubin* and *ubiquinon*, not indicating that *glutathion* and *bilirubin* are *hormones*. Some of these details will maybe be captured in an OELM that covers a greater amount of linguistic patterns and more data.

In general the OOF seems to contain more levels. Since it has been integrated with the SIMPLE-ontology, many of the concepts have been attached to already existing concepts. This means that some of the concepts have been attached to more detailed superordinate concepts or even more general ones depending on the concept in question. However, the resulting concept relations in the OELM seem to be adequate for an initial ontology. When terminologists work with domain modeling the result will always depend on means and purpose. In the present example, a nutrition expert should verify if the OELM is an adequate model or not. Surely an expert will complain and the ontology must then undergo corrections. Not only

is the ontology building process a very time consuming task in itself, it is also an iterative and dynamic process because it requires making corrections when discovering new terminological information.

5. Conclusion and future work

In this paper I have presented the initial results of a Ph.D. project that aims at identifying semantic relations between concepts in Danish domain specific texts. Based on a small set of data, it has been possible to construct an initial ontology based on the information that was carried by the identified linguistic markers.

The next step will most certainly be to experiment by applying a statistical method to my training data in order to measure the effect. For that special purpose the naive Bayes classifier approach to Word Sense Disambiguation will be considered as a working method.

It is interesting to find linguistic markers in order to state semantic relations between concepts, and implemented as a linguistic tool, it could save terminologists a lot of time. One of my hypotheses is, however, that the linguistic patterns that indicate semantic relations could vary from domain to domain. Therefore the linguistic markers are to be tested in a corpus from a different domain. In order to test this hypothesis I assume that I will need a larger collection of linguistic markers, e.g. more verbs, as the markers presented in this paper are not very likely to vary across domains.

Another very interesting experiment would be to merge an automatic relation extraction module with the terminological tool for construction of concept systems, within the framework of the project CAOS (Madsen et al., 2002). Since the aim of CAOS is to aid the terminologist in organizing and controlling the concepts and their relational features, this merging would probably facilitate the process even more.

6. References

- Jurafsky D. & Martin J.H. 2000: *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, 2000.
- Madsen, Bodil Nistrup, Thomsen, Hanne Erdman & Vikner, Carl: *Computer Assisted Ontology Structuring*. In: *Melby, Alan (ed.): Proceedings of TKE '02 - Terminology and Knowledge Engineering, INRIA, Nancy, 2002, s. 77-82.*
- TERMINOLOGY WORK — VOCABULARY — Part 1: Theory and application (Partial revision of ISO 1087:1990)
- Weilgaard, Lotte 2002. "Danish Verbs as Knowledge Probes in Corpus-based Terminology Work". In: *LSP & Professional Communication, Volume 2, Number 2, October pp. 77-93.*
- WordSmith <http://www.liv.ac.uk/~ms2928/>