Towards ontology based search and knowledge sharing using domain ontologies

Sine Zambach (sz@ruc.dk), Roskilde University, CBIT

November 11, 2008

Abstract

This paper reports on work in progress. We present work on domain specific verbs and their role as relations in domain ontologies. The domain ontology which is in focus for our research is modeled in cooperation with the Danish biotech company Novo Nordic. Two of the main purposes of domain ontologies for enterprises are as background for search and knowledge sharing used for e.g. multi lingual product development. Our aim is to use linguistic methods and logic to construct consistent ontologies that can be used in both a search perspective and as knowledge sharing. This focuses on identifying verbs for relations in the ontology modeling. For this work we use frequency lists from a biomedical text corpus of different genres as well as a study of the relations used in other biomedical text mining tools. In addition, we discuss how these relations can be used in broarder perspective.

Introduction

This work is a part of the SIABO project, a coorperation between the Danish universities Roskilde University, Danish Technical University, Copenhagen Business School and Novo Nordic.

The long term main scope of the work of our project, the SIABO-project, is to build a domain specific ontology based search engine [4] for information retrieval of domain specific text. The texts can for example be research notes, scientific literature or patents.

In this project we have started out with investigating the scientific literature and patents on the main area of Novo Nordic, namely concepts concerning diabetes. A part of this work, which is the focus of this paper, is to perform domain analysis for ontology modeling.

Relations and especially verb relations have been well studied over time and are integrated in many formalism. The semantic relations are important for the expressibility of a system since they glue the concepts together and decides what accosiations can be made as well as semantic types for the concepts around it. An example is:

"Glycose *stimulates* insulin secretion". The verb *stimulates*, is a positive regulatory relation that indicates that the noun or noun phrase before (in this case glycose)

If only *is a* (class subsumption) hierarchies or simple glossaries is used, the system cannot easily access associated terms that are not direct sub- or super classes. However, if all verbs related to the domain area is mapped to seperate relations, the ontology will be too heavy and intractable. In addition the information will be hidden if every verb corresponds to a relation and there will not be any gain of having verbs with same semantics clustered. Thus we need to categorize (and organize) the verbs in groups and focus on those verbs important for our goal.

Other groups have been working on ontologies for textmining in biomedical texts and many of them are using relations as a central part of their mining process [7, 2, 5].

In our work verb relation identification in texts will be used for:

- logic and reasoning between classes as "proof reading" for ontologies
- pattern extraction of the phrases that surrounds them. Described in details in the "Results and Discussion" section.

In a broader context the ontologies (using the identified relations) can be used for example for search and querying through scientific notes and patents (etc), and clarification of the connections between concepts in a domain language used e. g. for translation and consistant naming of software products (knowledge sharing).

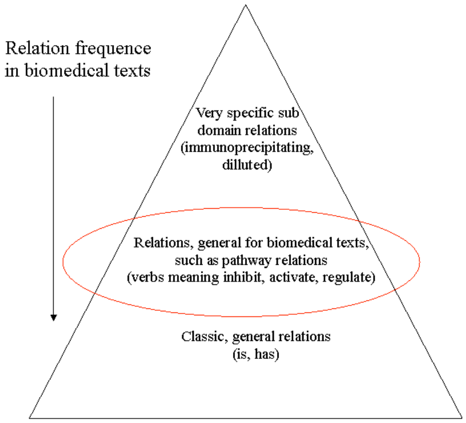
Our main focus is the biomedical industry – however the methods and ideas can be used in other enterprises and akademic areas as well. In this paper we will present the domain analysis using verb frequencies of four document classes:

- Patents concerning diabetes and stem cells [8]
- Medline abstracts [9]
- BioMed abstracts [1]
- British National Corpus (BNC, common British language) [6]

Verb frequencies and relations

When choosing relations we need to identify relations that are specific for the biomedical area, but still not so specific that they cannot be used within other sub areas of molecular biology than diabetes. Thus very specific verbs like methods for microbiological lab work (e. g. *immunoprecipitating*, *diluting*) was not considered. Some other text mining approaches [2, 7, 5] has been focusing on the pathway relations - central relations that connects substances in biomedical texts. These are typical positive (a *activates* b) or negative (a *inhibits* b). The

Figure 1: Graphic overview of the frequence of verbs representing relations in biomedical texts. Some relations are general to all common language domains, some are specific and differ among different research area and some are general for biomedical texts.


Relation frequence
in biomedical texts
Very specific sub
domain relations
(immunoprecipitating,
dilluted)
Relations, general for biomedical texts,
such as pathway relations
(verbs meaning inhibit, activate, regulate)
Classic, general relations
(is, has)

positive relation has the property of transitivity, whereas the negative one is more complex, though it still has a transitive-like behaviour.

In addition to these domain specific verbs, the general commonly used and well studied relations represented by the verb phrases such as *is a* and *has (the part)* will be used in the final ontology as well.

Being a Bioinformatician rather than a Computer Linguist, my aim is to analyze the domain and point out the relations that are widely used within biology and that can be grouped into few semantic meanings. Figure 1 illustrates roughly the frequency of different kinds of terms in biomedical texts. Some are very specific for several methods (e. g. *dilute*, *immunoprecipitating*) some are general, well studied and highly frequent in many different texts (*is* in combination with *a*, *part of* etc.) [3, 10].

In between is the frequent relations mentioned above that appears in most biomedical texts (*activate*, *inhibit* etc.). These are highly represented in the frequency list of both patent verbs concerning biomedicine and a collection of biomedical texts when compared to the common language text. In addition to our own experiments, other search tools concerning search in biomedical literature has identified similar relation groups [2, 7, 5]. Thus, we chose to start with focusing on the stimulatory and inhibitory relations also proposed by others as important relations [2].

Results and discussions

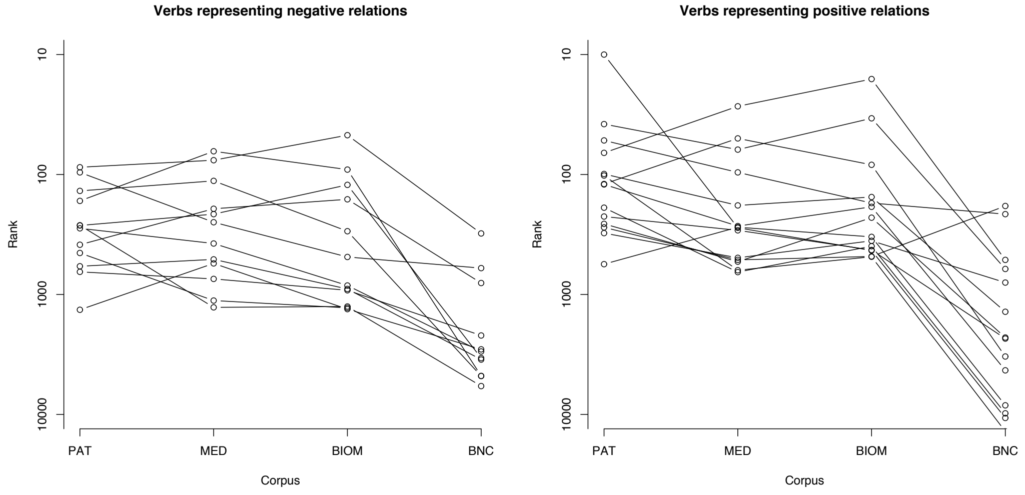
The selected verbs and their frequencies can be found in table 1 [11] and figure 2. This indicates a couple of trends:

- It shows that most biomedical verbs has a much higher ranking (is much more frequent) in biomedical texts than the common language corpus (BNC) which could indicate that the verbs contains more information than other more common verbs and that they could be object for relations to be used by the developer when building the background ontology.

- Words that are common in normal texts (*start*) but have a specific meaning in biomedical context (*stimulate*) has a much higher ranking in the BNC corpus than in the bio specific ones.

- It is not indifferent what kind of domain text we analyse. The patents, though they only represent a narrow part of the biomedical area still have some differences in biomedical word frequencies compared to Medline and BioMed. For example the words *encode, inactivate* and *remove* has a very low rank in the patents. A qualified guess for this issue is that many patent authors write their texts in a cryptic general or non-informative way to make it more difficult to get information and thus make illegal copies, or competitors can gain important knowledge that you want to hide etc.. The more typical biomedical verbs like *increase, induce* and *decrease* are on the other hand less frequent in the patent texts. (which the diabetes/stem cell focus in the patent texts can not explain).

Figure 2: A plot of the ranks for each verb in the four corpus from table 1. To eliminate insignificant differences we have used an inverse logaritmic scale. Although patent (PAT) verbs differ, it is the BNC-corpus that is the most different one. In contrast, Medline(MED) and BioMed(BIOM) are very similar.



- The pattents, Medline and BioMed are still much more similar to each other than the BNC-corpus, which is illustrated in figure 2.

Another interesting indication of the frequency lists are that the verbs identified by the Chilibot project [2] has a pretty high rank, which might mean that they are not used very frequently in neither the patent texts nor the BNC. However, these verbs has a bit lower ranking in the Medline and BioMed corpus, indicating that they are somehow more frequent here though still not widely used.

A question for discussion is how this work on verbs and relations can be integrated in a larger knowledge model for utilization of industries or akademia. We ourselves has the two purposes of the work, namely to model a background ontology and to be able to map different text corpus into the ontology by indexing and perhaps use automatic ontology generation from the relations. However the perspective might be wider than getting a better search through different kinds of literature.

The resulting ontologies can theoreticaly also be used for prediction of text corpus types using problem based methods as Hidden Markov Models, Neural Networks or for reasoning through the ontology.

Future work

The future goal is to develop the search system of Novo Nordic or similar enterprises. This implies to:

- Implement the relations in the ontology modeling
- Use the relations for knowledge pattern extraction. Knowledge patterns can be used to identify texts-fragments containing verbs that represent the relations. The fragments can be tagged with ontological types identified by patterns involving for instance noun phrases (NPs) as well as verbs. These ontological types on the NP's on each side of the relations can then be mapped into the ontology. Practically in system development, this mapping is similar to indexing if one uses a database to represent the data.
- Together, the ontology that we model and the indexing can end up in a *domain ontology based search engine.*

In the focus of this domain analysis work, which is a part of the system development mentioned above, the work must be refined with the following actions:

- At first we would like to analyze the whole BioMed corpus - not just the abstracts (since they, in contrast to the Medline articles, are open access we should take advantage of this).
- We should make an extensive concordance analysis to verify or falsify that the verbs actually carry the semantics of the relations in the biomedical texts in most cases.
- We could do a more qualitative investigation on how biomedical researchers actually understand the words, and if they have some preferences that we have not discovered through our initial qualitative work..

References

[1] BioMed Central. Publisher of 190 peer-reviewed open access journals. *http://www.biomedcentral.com/*, 2008.

[2] Hao Chen and Burt M Sharp. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 5(NIL):147, 2004.

[3] D A Cruse. On the transitivity of the part-whole relation. *Journal of Linguistics*, (15):1–201, 1979.

[4] T Andreasen et al. Ontological Extraction of Content for Text Quering. *NLDB*, LNCS 2553:123–136, 2002.

[5] R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke, and A. Valencia. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE*, 2005:pe21, May 2005.

Table 1: Verb ranks in Biomedical patents, Abstracts from BioMed Central, Medline and the British National Corpus (BNC). R is short hand for "rank". When more forms was present the lowest rank is used.

Verb	Semantic relation	R (patents)	R (Medline)	R (BioMed)	R (BNC)	Source
reduce	negative (inhibit)	87	76	47	310	[11]
remove	negative (inhibit)	96	250	487	603	[11]
inhibit	negative (inhibit)	137	113	297	4789	[11]
decrease	negative (inhibit)	166	64	91	4778	[11]
delete	negative (inhibit)	265	1281	1260	5806	[11]
regulate	negative (inhibit)	266	214	122	3391	[11]
block	negative (inhibit)	281	376	842	2987	[11]
limit	negative (inhibit)	385	193	161	803	[11]
inactivate	negative (inhibit)	451	1124	1293	NA	[11]
suppress	negative (inhibit)	582	510	900	3490	[2, 11]
eliminate	negative (inhibit)	648	742	923	2198	[2, 11]
attenuate	negative (inhibit)	915	736	796	17368	[2, 11]
abolish	negative (inhibit)	1342	551	1325	2868	[2, 11]
encode	positive (stimulate)	10	278	430	10704	[11]
express	positive (stimulate)	38	62	34	613	[11]
produce	positive (stimulate)	52	96	173	214	[11]
increase	positive (stimulate)	66	27	16	515	[11]
generate	positive (stimulate)	99	181	154	1394	[11]
secrete	positive (stimulate)	102	628	485	13853	[11]
induce	positive (stimulate)	120	50	83	3290	[11]
activate	positive (stimulate)	121	269	186	4291	[11]
amplify	positive (stimulate)	189	651	396	9795	[11]
stimulate	positive (stimulate)	224	292	426	2333	[11]
start	positive (stimulate)	276	515	482	183	[11]
promote	positive (stimulate)	308	495	357	796	[11]
facilitate	positive (stimulate)	258	533	229	2285	[2, 11]
elevate	positive (stimulate)	559	275	331	8400	[2, 11]

[6] Adam Kilgarriff. Assorted frequency lists and related documentation for the british national corpus (bnc). *http://www.kilgarriff.co.uk/bnc-readme.html*, 1995.

[7] Hans-Michael Muller, Eimear E Kenny, and Paul W Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, 2004.

[8] Diabetes patents. List of patents related to diabetes provided by novo nordic. *Unpublihed*, 2007.

[9] Medline (Entrez Pubmed). Search data base for biomedical litterature. *http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed*, 2008.

[10] Barry Smith and Cornelius Rosse. The role foundational relations in the alignment of biomedical ontologies. *MEDINFO*, pages 444–448, 2004.

[11] Sine Zambach Tine Lassen. Verb Frequence lists, different corpus, unpublished. 2008.