Relational Annotation of Scientific Medical Corpora

A Case Study

Ann-Marie Eklund

Department of Swedish, University of Gothenburg, Sweden*

Abstract. In life science and biomedicine much knowledge resides as unstructured information in for instance bibliographic databases. To facilitate searching and categorisation of this information the database entries are annotated with terms or keywords, describing for instance diseases, treatments and anatomy. These annotations are limited to concept level and do not describe relations between terms, for example that a given treatment may be used for a given disease, even if this information is available in both the text and terminologies.
In this work we will present a possible approach to extend term annotations with relational information to add another dimension to concept focused annotation schemas. This approach could also be used to highlight implicit information and to structure knowledge.

1 Introduction

In life science and biomedicine much knowledge resides as unstructured information in for instance bibliographic repositories or electronic health record databases. To facilitate searching and categorisation of this information the database entries are annotated with terms or keywords, describing for instance diseases, treatments and anatomy. These annotations are limited to concept level and do not describe relations between terms, for example that a given treatment may be used for a given disease, even if this information is available in the paper. This annotation limitation remains in spite of accessible relational information in ontologies and terminologies.

In this work we study the possibility to annotate a term annotated scientific medical corpus with relations between the terms, for instance relating diseases to treatments and organ sites. We use a Swedish text corpus of scientific medical documents [3], which has been annotated with information from the medical terminology MeSH (Medical Subject Headings)[1].

Current related work has focused on for instance methods for extraction of specific types of relations [5,6] and extraction and characterisation of semantic

* The author would like to thank Centre for Language Technology, Gothenburg (clt.gu.se) for financial support.

[1] www.nlm.nih.gov/mesh

relations [1,2,4] in biomedical text. For instance, Abacha and Zweigenbaum [1], Frunza and Inkpen [2] and Yao et al [6] studied cure, prevent and side effect relations in medical papers and Segura-Bedmar et al [5] studied resolving anaphoras for extraction of drug-drug interaction in pharmacological documents.

However, our work focuses not on establishing methods for extraction of relations between terms, but on tying existing term annotations together, reflecting relations in the text. These new relational annotations would allow both medical scientists and search engines to take advantage of highlighted implicit information on for instance diseases, treatments and anatomy. For example a paper regarding prevention of myocardial infarction may be annotated with terms like Aspirin and Myocardial Infarction and our proposal is to also annotate it with the relation *may_prevent* relating these terms. Not only would this add new useful annotations, but it would also structure existing annotations.

2 Materials and Methods

2.1 Materials

The main resources in this study are an annotated Swedish scientific medical corpus [3] and the vocabularies and terminologies of the Unified Medical Language System[2].

Medical Text Corpus. As a part of the Swedish strategy for e-health, the clinical terminology SNOMED CT[3] has been translated into Swedish. For validation and quality assessment of the translation, a Swedish medical text corpus was created from the electronic archives of the Journal of the Swedish Medical Association 1996-2009 [3]. The corpus comprises 29110 documents (28 million tokens) and has been part-of-speech tagged and annotated with Swedish and English MeSH (release 2006) and with the Swedish SNOMED CT.

Our study focuses on the MeSH annotated sentences in a part of the corpus containing 2021 articles from the domain "Klinik och Vetenskap" (Medical Practice and Science). This part contains 140458 sentences, each with a unique id. For copyright reasons the order of the sentences have been randomised, thereby limiting our study to relational annotations at sentence level. Of this sentence set we used only the 102821 sentences with at least two MeSH annotations.

UMLS. The Unified Medical Language System (UMLS)[4] connects vocabularies from different biomedical and health-related sources in different languages. It provides, among other things, databases, called Knowledge Sources. One of the databases, Metathesaurus, contains information about more than one million biomedical or health-related concepts. This database is divided into a number

[2] www.nlm.nih.gov/research/umls
[3] www.ihtsdo.org
[4] We have used UMLS version 2010AB, including all source vocabularies of level 0-3.

of relational tables. One of the major tables, MRCONSO, contains the structure for each concept, e.g. names, identifiers, languages and source vocabularies. This table is complemented with the MRSAT connecting MeSH identifiers (or identifiers from other source vocabularies) to concept identifiers (CUI).

Metathesaurus also contains relations between concepts, where the table MRREL contains basic relations (REL), e.g. Parent/Child, relating different concepts. Around 25% of the relations have a label (RELA - Relationship Attribute) which comes from the source vocabulary and specifies the relationship, e.g. *isa*, *treated_by*, *finding_site_of* or *has_component*.

Considering the MeSH part of the UMLS, a term can belong to more than one category and thereby appear in several places in the MeSH hierarchy with different MeSH identifiers. Moreover, by the annotation procedure used in our study corpus, a term like Blood Pressure will be annoted as Blood Pressure, but also as Blood and Pressure. Hence, the MeSH annotations can be nested, which can result in more UMLS concepts than found terms in a sentence.

2.2 Methods

Since this work can be seen as a feasibility study of extending term annotation schemas with relational information, the methods have been kept simple and analysis and validation of the approach were done by manual inspection.

For each of the sentences in our corpus containing at least two MeSH annotations, we extracted the sentence identifier and MeSH annotations, not taking into account any nesting of the annotations. The resulting information was stored in a MySQL database, complementing the UMLS one, thereby allowing easy mapping of MeSH identifiers to UMLS concepts via the MRSAT table. These concepts could then be used to derive relations from MRREL, giving relations between the annotated terms in each sentence.

Since our main interest is in the ability to provide relational annotations among annotated terms, the analysis focused on the derived relations with information in the RELA field of MRREL. One such RELA relation is *may_prevent*, e.g. "Aspirin *may_prevent* Myocardial Infarction". These relations were divided into five different categories reflecting disease-treatment, disease-organ, cause-effect, hierarchical and other relations among the terms to reflect relations often found in medical papers. The division of the RELA relations into these classes was based on our subjective interpretation of relations like e.g. *may_treat*, *has_finding_site* and *cause_of*. For each of the relation categories, we randomly picked sentences and manually compared the derived relations to the ones expressed in the sentences to see if and how they were related. Since this work is to be viewed as a feasibility study for future research, we limited our analysis to a handful of sentences per relational category.

3 Results

Our study corpus comprised 140458 sentences, with 102821 (73%) containing at least two MeSH annotations and in 26251 (25%) of these we were able to identify

relations between the terms. In these sentences we found 150024 relations of which 44754 (30%) had the specification RELA. There were 188 different RELA. Table 1 shows the percentage of sentences per RELA relation.

Table 1. Percentages of sentences (of 10854 sentences with RELA) per RELA relation.

Relation	Percentage
isa	37
co-occurs_with	10
associated_with	10
has_finding_site	6.8
sib_in_isa	6.8
mapped_from	6.0
location_of	5.3
is_associated_anatomic_site_of	4.0
may_treat	2.8
part_of	2.4
may_prevent	1.4
may_be_a	1.1
causative_agent_of	0.8

Figure 1 shows the 122483 sentences containing MeSH annotated terms divided into number of MeSH terms (left) and number of sentences per number of relations (right)[5].

In the rest of this section we analyse each of the defined relation categories and exemplify relations and their corresponding sentences, Table 2, and in Discussion we briefly elaborate on these results.

Disease - Treatment – The relations between diseases and treatments found in the studied part of the corpus are e.g. *may_treat* and *may_prevent*. Examples of disease-treatment relations are the relations derived from sentences 77591, 50380 and 105057. From sentence 105057 it is not possible to infer what the relation is between the terms Soft Tissue Infections and Methicillin.

Disease - Organ – There are a number of different relations between diseases and organs, e.g. *is_associated_anatomic_site_of* and *has_finding_site*. The relation *location_of* is often a relation between diseases and treatments, but sometimes refer to organ-organ relations. Disease-organ relations were found in for example sentences 75743, 8039 and 97183. In 75743 and 8039 the relation was not explicitly expressed.

Cause - Effect – Cause and effect relations can be for example *cause_of*, *induces* and *causative_agent_of*. It can be for instance viruses or bacteria which cause diseases, or diseases that cause other diseases. Examples of these relations were found in sentences 133130 and 125988.

[5] In the UMLS, for many relations there is also an inverse, e.g. *finding_site_of* and *has_finding_site*.

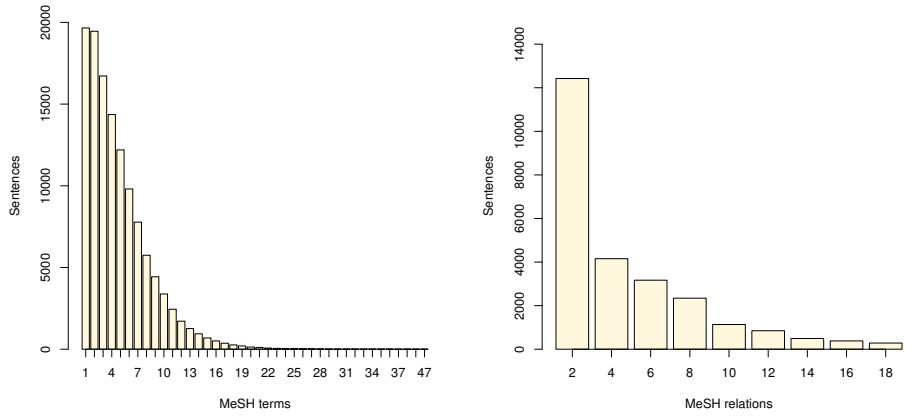

Fig. 1. Number of sentences per number of MeSH terms (left) and number of sentences per number of relations (right).

Hierarchical Relations – The hierarchical relations we have studied are synonymy, hyponymy and sibling relations. Synonyms are for instance concepts which have the relation *same_as* or *has_tradename*. The relations *mapped_from* and *primary_mapped_from* can be synonym relations, sentence 77314. Hyponymy can be relations like *isa*, *may_be_a* and *part_of*, sentence 43382. Sibling relations found in the sentences are for instance *sib_in_isa*. In the example of the sibling relation, where Ethanols, Methanols and Ethylene Glycols are all Alcohols (sentence 14668), the relations have no specification in RELA, only the REL abbreviation SIB.

Other Relations – Other relations that were found in the sentences in the corpus are for instance *co-occurs_with*, *associated_with*, *may_diagnose* and *occurs_before*.

Sentences 15403 and 12826 are examples of the relation *co-occurs_with*, while sentences 52963, 51136 and 125709 are examples of the relations *associated_with*, *may_diagnose* and *occurs_before* respectively.

4 Discussion

The examples in this work show that it is possible to derive relations from the annotated terms in a sentence utilising only UMLS. The majority of the examples for the different relation types are existing relations, but sometimes the derived relations are implicit instead of directly expressed in the sentences. However, as exemplified in sentence 75743, the derived relations may not be the ones intended in the sentence.

Since we have studied only concepts and not taken into account the syntax of the sentences, the resulting relations between the terms can be relations not

Table 2. Derived relations and sentences (English translations).

Sentence ID	Relation
77591	Opium *may_treat* Pain
50380	Aspirin *may_prevent* Myocardial Infarction
105057	Soft Tissue Infections *may_be_treated_by* Methicillin
75743	Pancreas *is_associated_anatomic_site_of* Diabetes Mellitus
8039	Celiak Disease *has_finding_site* Intestines
97183	Central Nervous System *finding_site_of* Rabies
133130	Bacillus antracis *causative_agent_of* Anthrax
125988	Nitrous Oxide *induces* Nausea
77314	Arthritis *primary_mapped_from* Joint Diseases
43382	Estrogenes *isa* Hormones
14668	*SIB*: Ethanols/Methanols/Ethylene Glycols
15403	Obesity *co-occurs_with* Diabetes Mellitus
12826	Diabetes Mellitus *co-occurs_with* Hypertension
52963	Vaccination *associated_with* Immunization
51136	Triiodothyronine *may_diagnose* Thyroid Disease
125709	Chickenpox *occurs_before* Herpes Zoster (shingles)

Sentence ID	Sentence
77591	[...] opium had a soothing effect on both anxiety and pain.
50380	[...] observed that patients with a regular intake of Aspirin had fewer heart attacks than expected.
105057	[...] soft tissue infections [...] and infections caused by methicillin resistant staphylococcus.
75743	An alternative for a few patients with diabetes mellitus has been transplantation of pancreas [...]
8039	[...] patients with already known celiak disease but who in spite of a strict diet have had gastro-intestinal symptoms.
97183	[...] the symptoms of a rabies infection begins when the virus reaches CNS [...]
133130	Bacillus antracis causes the disease Anthrax [...]
125988	[...] nitrous oxide contributes to post-operative nausea.
77314	[...] patients with [...] and arthritis symptoms who were treated with [...] improvement of their joint disease [...]
43382	Even though it is well documented that estrogene [...] is an important hormone in [...]
14668	[...] have been introduced as an alternative to ethanol in cases of ethylene glycol and methanol poisoning.
15403	[...] obesity is associated with a highly increased risk of developing [...] diabetes [...]
12826	Other risk factors for stroke are [...] hypertension, diabetes, [...]
52963	Participation in the vaccination programs has been very high and sufficient immunization was reached [...]
51136	Antibodies targeting Triiodothyronine [...] in up to 10 percent of patients with Thyroid diseases.
125709	[...] a connection between chickenpox and herpes zoster.

derivable from a sentence. For instance, in sentence 105057 there is no indication of any relation between the terms Soft Tissue Infections and Methicillin, but from the UMLS we get a *may_be_treated_by* relation. Many of the derived UMLS relations are not explicitly expressed in the sentences, but can be part of the context of a sentence, for instance the relation between Diabetes Mellitus and Pancreas in sentence 75743.

Expressing hierarchical relations like *isa* and *part_of could* lead to increased understanding of the context of the sentence. For example in 14668, where, by the sibling relation, we learn that the terms Ethanols, Methanols and Ethylene Glycols have something in common, i.e. they are all Alcohols, thereby framing the concepts in the sentence.

The terms which have the relation *co-occurs_with* can have slightly different relations to each other in the sentences. For example in 15403 one problem leads to another, but in 12826 the two diseases can both be the cause of a third one.

One of the major reasons for only being able to identify relations in 25% of the sentences with more than two annotations, may be that the annotations are not at the same hierarchical ontological levels in comparison to the defined relations in UMLS. A fundamental challenge with the proposed approach is its dependence on the quality and source of the original term annotations, as in our case using only MeSH in the process to identify relational annotations.

5 Conclusions and Future Work

In this work we have studied the feasibility of utilising the term annotations of medical text in connection with a collection of terminologies and vocabularies to extend the annotations with relational information. As the examples show, this approach can be used not only for annotation, but also to highlight implicit information and to structure knowledge.

Our work complements existing research on extraction of (semantic) relations in biomedical text, by focusing on identifying and validating relations between conceptual annotations of a text. Hence, instead of the complex process of extracting relations in medical papers, we utilise existing annotations to propose potential relations covered by a paper.

This work is based on only one of the source vocabularies and a limited corpus. Hence, future work will address the ability to make use of combinations of several source vocabularies and more elaborated use of the hierarchical ontological relations to increase the ability to identify relations among annotated terms. However, utilising for instance hyponymy induces challenges like degree of hyponymy/hypernymy to allow in establishing relations among terms, and also how to resolve problems with the complex relational structure of the UMLS with many different types of relations and even cycles. For instance, some vocabularies in the UMLS may treat relations like *isa* and *part_of* as synonymous and some as distinct types of relations. Ongoing work also considers other corpora, like parts of MEDLINE[6], for relational annotation.

[6] www.nlm.nih.gov/databases/databases_medline.html

References

1. Abacha, A.B., Zweigenbaum, P.: A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In: Proceedings of 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011) (2011)
2. Frunza, O., Inkpen, D.: Extraction of disease-treatment semantic relations from biomedical sentences. In: BioNLP Workshop (ACL 2010) (2010)
3. Kokkinakis, D., Gerdin, U.: A Swedish scientific medical corpus for terminology management and linguistic exploration. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010) (2010)
4. Patel, C.O., Cimino, J.J.: Using semantic and structural properties of the unified medical language system to discover potential terminological relationships. J Am Med Inform Assoc 16(3), 346–353 (2009), http://dx.doi.org/10.1197/jamia.M2931
5. Segura-Bedmar, I., Crespo, M., de Pablo-Sanchez, C., Martinez, P.: Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. BMC Bioinformatics 11 Suppl 2, S1 (2010), http://dx.doi.org/10.1186/1471-2105-11-S2-S1
6. Yao, L., Sun, C., Wang, X., Wang, X.: Relationship extraction from biomedical literature using maximum entropy based on rich features. In: Proceedings International Conference on Machine Learning and Cybernetics, (ICMLC 2010). pp. 3358–3361 (2010)