Creating a Gold Standard Corpus for the Extraction of Chemistry-Disease Relations from Patent Texts

Antje Schlaf [1], **Claudia Bobach** [2], **Matthias Irmer** [2]
[1]Natural Language Processing Group
University of Leipzig, Germany
antje.schlaf@informatik.uni-leipzig.de

[2]OntoChem GmbH
Halle (Saale), Germany
{claudia.bobach, matthias.irmer}@ontochem.com

Abstract

This paper describes the creation of a gold standard for chemistry-disease relations in patent texts. We start with an automated annotation of named entities of the domains chemistry (e.g. "propranolol") and diseases (e.g. "hypertension") as well as of related domains like methods and substances. After that, domain-relevant relations between these entities, e.g. "propranolol treats hypertension", have been manually annotated. The corpus is intended to be suitable for developing and evaluating relation extraction methods. In addition, we present two reasoning methods of high precision for automatically extending the set of extracted relations. Chain reasoning provides a method to infer and integrate additional, indirectly expressed relations occurring in relation chains. Enumeration reasoning exploits the frequent occurrence of enumerations in patents and automatically derives additional relations. These two methods are applicable both for verifying and extending the manually annotated data as well as for potential improvements of automatic relation extraction.

Keywords: corpus creation, patent texts, relation extraction

1. Introduction

Patents provide a huge source of publicly available information and knowledge. For example, about 10 to 14 % of all patent applications deal with chemical compounds and their use in novel health or agricultural products. To extract this domain specific knowledge we are aiming to develop and apply automated knowledge extraction processes that are based on semantic named entity recognition (NER) as well as recognizing and extracting relevant relationships between those named entities. Relationships between the chemical structure of a particular chemical compound and its biological activities (structure-activity-relationships or SAR) are of special interest for pharmaceutical or agricultural research. This SAR knowledge can then be used to predict novel compound properties or to design compounds with better properties.

For the present work our text mining focus was on pharmacological properties of chemical compounds, extracting knowledge triples in the form of <chemistry> <relation> <health condition> such as <propranolol> <treats> <hypertension> from patents.

The particular linguistic features of patents pose a challenge for any attempt to automatically extract information: sentences are often very long and may exhibit a high syntactic complexity when compared to other text documents (Verbene et al., 2010). As a consequence, state-of-the-art established statistical parsers are not very suitable for automated patent text processing as they often fail to successfully parse these long sentences. This failure becomes especially apparent for the highly complex syntax of patent claims (Parapatics and Dittenbach, 2011) that contain novel information on SAR knowledge. Thus, novel high performant semtantic text mining methods need to be implemented to achieve our goal of extracting SAR data.

For efficient information retrieval method development, a corpus of patent documents with manually annotated SAR data is of high interest. Such a gold standard corpus could also be used for evaluating and comparing methods and systems for the automated extraction of domain-relevant relations between entities mentioned in a patent text.

There are some related existing resources. In the ChiKEL project (Milward et al., 2012) and in the freely available ChEBI Patent Gold Standard[1], patent texts were annotated with chemistry terms but not with relations. PharmGKB[2] provides chemical relations, but no chemistry-disease relations. The Comparative Toxicogenomics Database[3] (CTD) contains chemistry-disease relations, though not for patents but for abstracts of research papers. Bartalesi Lenzi et al. (2009) provide a gold standard for relations in patents but in another domain (optical devices and machine tools). Thus, none of these resources provides data which can be used for the development of methods for extracting domain-relevant relations from patent texts. As a consequence, we decided to create a new, manually annotated gold standard corpus of patent documents.

This paper is structured as follows: Section 2 describes the creation of the corpus and points to specific challenges for the annotation of chemistry-disease relations in patent texts. Section 3 describes the exploitation of two automatic methods to enhance the gold standard creation process. Finally, we draw conclusions and point out possible directions for future steps in Section 4.

[1]http://chebi.cvs.sourceforge.net/viewvc/chebi/chapati/patentsGoldStandard/

[2]http://www.pharmgkb.org/

[3]http://ctdbase.org/

2. Creating a gold standard corpus: manual annotation of chemistry-disease relations in patents

2.1. The corpus

For creating the corpus, we randomly selected 21 US patent applications from 2010 which contained a claimed SAR relation. We annotated the named entities in these documents with OCMiner[4] and selected all of those sentences which contained both a chemistry as well as a disease term, resulting in a total of 365 sentences. This co-occurrence based selection method provides a first approximation to the maximal recall of chemistry-disease relations expressed in the text. Please note that relations that are expressed by more than one sentence were not considered for the present corpus. Subsequently, sentences were manually annotated with the help of the open-source tool BRAT[5]. A screenshot of a BRAT annotated sentence is depicted in Figure 1.

2.2. Named entities

Recognition of named entities (NEs) was performed by OCMiner, a high-performance text processing system based on Apache UIMA[6] using dictionaries created from domain-specific ontologies (e.g. a chemistry ontology, Bobach et al., 2012). Table 1 shows the different named entity types used in this work. For example, NE type "chemCmpd" refers to specific chemical compounds (e.g. "adenosine"), whereas "chemClass" refers to a concept for a family of chemical compounds – also described as compound classes in chemical terminology (e.g. "adenosines" or "adenosine derivatives", Irmer et al., 2013). Note that the rather unspecific expression "compound" is also annotated as a NE of type chemClass. This is especially relevant in patents as they often refer to a compound as "the [said] compound [of the present invention]", specifying the actual chemical compound elsewhere in the text. The NE type "anaphor" is used for annotating anaphoric terms like "their" or "thereof" which are coreferent to other entities mentioned before.

2.3. Relations

In this work, our focus lies on relationships between chemical entities and diseases with the main connection categories shown in Table 1. In principle, we did not consider other relations which might have been expressed in the same sentence. However, during the annotation process, we found particular difficulties for the annotation of SAR relations in patent texts. A series of consequences emerged from these observations.

Distances between relata may be very large. This is especially true for sentences in patent claims. Furthermore, the relationship between chemistry and disease terms is often not directly expressed, but rather via relation chains, for example "[compound] is part of [substance] is applied in [method] for treatment of [disease]", which can be seen in the example patent sentence in Figure 2. Thus, we decided to annotate relations in a fine-grained way in order to keep a close proximity to the text. This leads to easier annotation rules and shorter distances between relata. It prevents human annotators from inferring indirect relations based on their world knowledge but not expressed in the text. Furthermore, such a fine-grained gold standard is qualified for the development of methods for the automatic relation extraction from patent texts. The relations indirectly expressed via relation chains can automatically be retrieved with the help of chain reasoning described in Section 3.1.

Named entity types	Main connection categories	Further connection categories
chemCmpd, chemClass, disease, method, substance, anaphor	treats, induces, doesNotInduce, resistance, modulates, relatesTo	isInstanceOf, isActivePartOf, coreference, fusion

Table 1: Named entity types and connection categories used for annotation in the corpus.

In addition to the main connection categories mentioned above, further specific connection categories are useful:

isActivePartOf describes the fact that an entity is an active part of another entity. For instance, a chemical entity can be the biologically active ingredient of a substance, which, in turn, may be an active part of a method.

isInstanceOf has been used to connect entities of the same domain and different specificity. This is used in formulations such as "... for the treatment of sex hormone dependent diseases such as prostate cancer, breast cancer ..." where "prostate cancer" is an instance of the more general concept "diseases".

coreference is a relationship used to indicate relations between anaphoric terms and their antecedents. This is especially useful if anaphora form a part of our target relations.

fusion is a relationship that permits to infer composite entities formed by single named entities. The reason for introducing this special kind of annotation is that these single entities may contain anaphoric terms, whereas the actual entities can be spread over wide distances and may contain further sub-enumerations (see Figure 1).

2.4. Annotation results

In a first round of annotations, 875 relations were manually annotated in 145 sentences. They contain relations of different combinations of named entity types and connection categories as defined in Table 1. For example, Table 3 shows the number of annotated relations of type R for <chemClass> <R> <disease> and <chemCmpd> <R> <disease>. As shown by these examples, only few relation types were found between chemical and disease terms, especially for named entities of the type "chemCmpd".

In the following section, we propose a method for automatically inferring these specific relations via relation chains. The implementation of this method shall permit a rapid expansion of the gold corpus.

[4] http://www.ocminer.com/

[5] http://brat.nlplab.org/

[6] http://uima.apache.org/

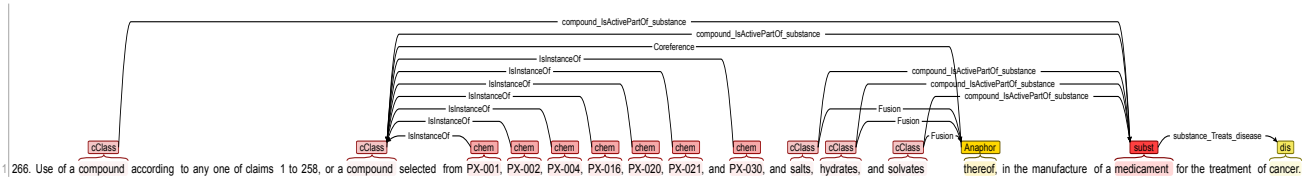

Figure 1: Example sentence visualizing the "anaphor" entity type and the "fusion" connection category (screenshot of the BRAT-Tool).

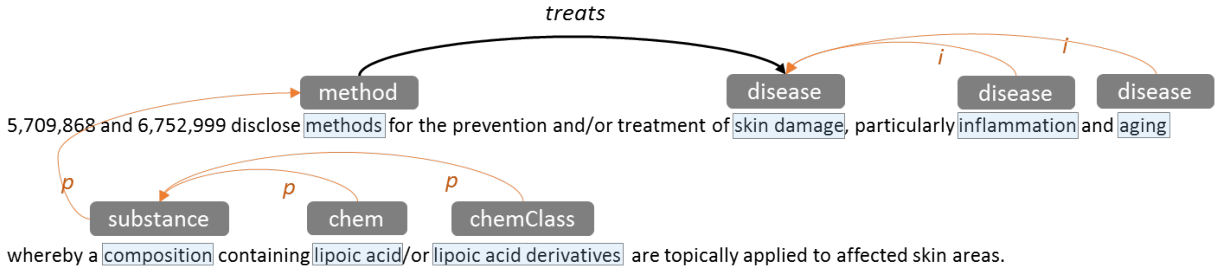

Figure 2: Example of relations in a patent sentence. The black arrow shows the relation of the main connection category. The orange arrows show further relations not of main interest, but attached to the relata of the main relation, where p=isActivePartOf, i=isInstanceOf.

3. Automatic improvements in the gold standard creation process

3.1. Chain reasoning

Chemical entities and disease terms are often connected via chains represented by additional relations appending on a main relation (see example sentence in Figure 1). Chain reasoning may provide a method to infer and integrate those additional, indirectly expressed relations. Thus, for the implementation of chain reasoning, we are proposing to distinguish between main relations, having a main connection category (see Table 1), and appending relations, with the connection category "isInstanceOf", "isActivePartOf" or "coreference". Chain reasoning was implemented as follows: If one of the relata of a main relation has a connection to another named entity via an appending relation, the new relation is inferred by replacing the original relatum by the appending named entity. For example:

Given relations: <method> <treats> <skinDamage>,
<skinAging> <isInstanceOf> <skinDamage>

Resulting reasoned relation:
<method> <treats> <skinAging>

However, this approach may not be valid in all cases. Therefore, we have applied specific rules for defining cases where chain reasoning is allowed within the context of a sentence (see Table 2 and Figure 3).

Since the reasoned relations are main relations, we can use them to perform a further reasoning on them by inspecting their appending relations and again applying the reasoning rules above. Thus, we have an iterative process of reasoning until no more new relations are inferred. Using this method, a total of 1397 new relations were inferred after 3 iterations.

appending relations	O–L	O–R	I–L	I–R
isInstanceOf	–	–	X	X
isActivePartOf	X	–	X	–
Coreference	X	X	X	X

Table 2: Inferring rules: An X marks reasoning for the given configuration as allowed.

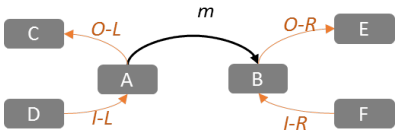

Figure 3: All possible directions of appending relations attached to the relata of a main relation m. O–L = Outgoing from left relatum, O–R = outgoing from right relatum, I–L = incoming to left relatum, I–R = incoming to right relatum.

The last column of Table 3 shows the number of chemistry-disease relations after chain reasoning.

The success of chain reasoning is demonstrated by a tremendous increase of new relations of <chemClass> <treats> <disease>, <chemCmpd> <treats> <disease>, and <chemCmpd> <relatesTo> <disease>. For other main relations no or only few additional relations were found. One reason for this finding might be that most likely more patents report about treating diseases with compounds rather than inducing diseases with compounds. To estimate the precision of chain reasoning for our defined rules as well as for the iterative reasoning, we randomly selected a subset of 400 relations found by the chain reasoning. 397 of these were manually evaluated as correct. The 3 relations

Pair	ConnectionCategory	Manual annotation	After chain reasoning
chemClass-disease (1008)	noRelation	890	301
	doesNotInduce	2	2
	induces	10	11
	relatesTo	5	5
	treats	101	**689**
chemCmp-disease (713)	noRelation	678	179
	doesNotInduce	2	2
	induces	3	9
	modulates	2	2
	relatesTo	15	**217**
	resistanceOf	8	15
	treats	5	**289**

Table 3: Number of relations between chemical named entities (chemClass/chemCmpd) and diseases in manual annotation and after chain reasoning.

marked as wrong were caused by errors in the manual annotations. Though we did not evaluate the exact recall due to the huge effort, we conclude that chain reasoning retrieved a huge amount of new hidden relations with very high precision, most importantly for the most specific and interesting named entities of type "chemCmpd".

The application of chain reasoning is twofold: On the one hand, the presented chain reasoning method can be used to automatically expand the number of annotated relations by inferring indirect relations, enabling the generation of an exhaustive set of high quality relations. On the other hand, automatic relation extraction might have higher quality if, in a first step, only direct relations are retrieved. As a second step, chain reasoning can be used to retrieve indirect relations.

3.2. Enumeration reasoning

Patents often try to cover as much fields as possible and therefore often contain enumerations for example by lists of named entities that all could be part of <chemistry> <treats> <disease> knowledge triples (see Figure 4).

Our hypothesis is as follows: If there is a valid relation $<R_1><X><R_2>$ with R_2 being located in an enumeration, the probability should be very high that $<R_1><X><R_i>$ is a valid relation as well for all R_i located in the enumeration together with R_2. Thus, $<R_1><X><R_i>$ can be set as a valid relation in enumeration reasoning. To validate our thesis we implemented a simple, prototypical enumeration identification: Named entities co-occurring in a sentence are defined as an enumeration if there is no verb between them and if they are all of the same type and no named entity of another type occurs in between (see description of named entity types in Table 1). We performed enumeration reasoning as described above on disease enumerations and all pairs of <chemCmpd> <X> <disease> and <chemClass> <X> <disease> relations in the gold standard including chain reasoning. Our thesis was proven correct for the data used: 98.95 % (1041/1052) of the reasoned relations were gold relations and therefore correct. The manual annotators found 2 missing relations in the 11 relations not being in the gold standard, so the actual rate was increased to 99.14 %. Other errors were caused by our implemented method to identify enumerations, which might be further improved in the future. As a result, enumeration reasoning exhibits a very high precision and can therefore be used to correct manual annotation errors. Moreover, time and effort can be saved if annotators only have to assign one entity within an enumeration into a relation, while the remainder can be inferred automatically. Enumeration reasoning may also help in automatic relation extraction systems to increase recall with relations of very high precision.

Specifically, the compounds of this invention are also useful in the treatment of various CNS disorders, hematological disorders, eating disorders, in the treatment for pain, respiratory diseases, genito-urological disorders, cardiovascular diseases and the treatment of cancer.

Figure 4: Example of a patent sentence containing an enumeration of diseases.

4. Conclusions and future work

We created a gold standard corpus for chemistry-disease relations in patents. We developed methods for manual annotation and defined required named entity types and connection categories necessary for specifying these relations within the complex structure of patent sentences. Chain reasoning was introduced to automatically generate an exhaustive set of very high quality gold relations. The method can also be used to retrieve indirect relations after an automatic relation extraction. Enumeration reasoning exploits the frequent occurrence of enumerations in patents and automatically derives relations with very high quality. In addition, it can also be used to find and correct manual annotation errors, to save time and manual annotation efforts and to boost the recall of automatic relation extraction processes with high precision relations.

Future work will include further extension of the gold standard corpus by manually annotating more sentences, as well as by the annotation of patent full texts, or at least, entire claim sections. Furthermore, we plan to create manually an-

notated data for enumerations in order to have specific evaluation data for the presented approach as well as for more complex approaches to enumeration identification in this domain. Last but not least we will use the created gold standard to evaluate methods for automatic relation annotation.

5. Acknowledgement

This work was funded by a grant of the German Federal Ministry of Education and Research (BMBF), project "SARminer" (01IS12011).

6. References

Bartalesi Lenzi, V., Sprugnoli, R. and Pianta, E. (2009): Annotation of Semantic Relations in Patent Documents. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon* (GL2009), Pisa, Italy.

Bobach, C., Böhme, T., Laube, U., Püschel, A. and Weber, L. (2012): Automated compound classification using a chemical ontology. *Journal of Cheminformatics* 4(1), 40.

Irmer, M., Bobach, C., Böhme, T., Püschel, A. and Weber, L. (2013): Using a chemical ontology for detecting and classifying chemical terms mentioned in texts. In *Proceedings of Bio-Ontologies 2013*, Berlin.

Milward, D., Corbett, P. and Bonnoit de Ruisselet, D. (2012): *Advances in text mining for chemical information: an update on the ChiKEL Project*. http://www.chemaxon.com/library/us-ugm-2012/advances-in-text-mining-for-chemical-information-an- update-on-the-chikel-project/

Parapatics, P. and Dittenbach, M. (2011): Patent claim decomposition for improved information extraction. In Lupu, M., Mayer, K., Tait, J., Trippe, A. J., and Croft, W. B. (Eds), *Current Challenges in Patent Information Retrieval*, 197–216. Springer, Berlin/Heidelberg.

Verberne, S., Koster C. and Oostdijk, N. (2010): Quantifying the challenges in parsing patent claims. In *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval* (AsPIRe 2010), 14–21.