OCMiner for Patents. Extracting Chemical Information from Patent Texts

Matthias Irmer*[1], Lutz Weber[1], Timo Böhme[1], Anett Püschel[1], Claudia Bobach[1], Ulf Laube[1]

OntoChem IT Solutions GmbH
Halle (Saale), Germany

[1]<firstname>.<lastname>@ontochem.com

Abstract. This paper describes OCMiner, a high-performance semantic text processing system for large document collections of scientific publications, and its performance regarding chemical named entity recognition in patent texts within the BioCreative V CHEMDNER-Patents challenge which was set up for this purpose. OCMiner permits adjusting the quality of annotation results by several linguistic options, which can be specialized and fine-tuned for the recognition of chemical and other Life Science terms. Recognized terms are mapped to semantic concepts which are ontologically located within their respective domain taxonomies. If possible, a chemical structure is assigned to chemical compound expression. Annotated document collections, among them US patent applications and grants, can be visualized on a web-based front-end at http://www.ocminer.com/.

Keywords. Chemical Named Entity Recognition, Patent Text Mining, Concept Mapping

1 Introduction

Patents provide a huge and steadily growing source of publicly available information and knowledge. For example, up to 14% of all patent applications deal with chemical compounds and their use in novel pharmaceutical or agricultural products. To extract this domain specific knowledge we are aiming to develop and apply automated knowledge extraction processes. The quality of such extracted information relies on a correct named entity recognition (NER) of chemical compounds mentioned in patent texts. This recognition process represents a particular challenge due to a variety of reasons: First, the peculiar linguistic features of patent texts 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 [1]. Second, there is a great variety in which chemical entities are referred to in texts: For example, there are both trivial, half trivial and systematic names for chemical compounds and classes, as well as formulas, registration numbers and trade names for drugs. Chemical names can be extremely long and may contain variations of meaningful punctuation symbols and parentheses. Different chemistry name types can even be mixed within one chemical expression. Further, patent texts are often available only as picture PDF documents from which the actual text has been extracted via error prone optical recognition (OCR) techniques. Especially standard OCR systems have particular difficulties in correctly recognizing numbers, parentheses and special characters within chemical expressions, e.g. "l" or "I" instead of "1", often resulting in misspelled chemical names. Third, patent authors generally tend to hide relevant information by using semantically underdetermined concept notations in connection with specifying and sometimes obscuring attributes. Specifically for chemistry patents, chemical compounds are often not explicitly mentioned but described by means of very complex and potentially nested Markush structure enumerations.

The BioCreative V CHEMDNER-Patents challenge [2] aimed at measuring the quality of recognizing mentions of chemical entities in patent texts. Two subtasks dealed with this question: the chemical entity mention (CEMP) task, which consisted in finding the exact textual positions of chemical entity mentions, and the chemical passage detection (CPD) task, which consisted in deciding whether a given textual passage (here: patent title or abstract) contains chemical entity mentions or not.

2 System Description

OCMiner is a modular processing pipeline for unstructured information based on the Apache UIMA framework [3]. Documents are read from a variety of sources (text and picture PDF, XML, etc.) and standardized for further analysis. Then, preparatory processes such as language detection, tokenization, document structuring, etc. take place.

As the core of the annotation process, we have a dictionary-based named entity recognition module which uses a high performance dictionary look-up technology with support for very large dictionaries. The dictionary module implements specific language and dictionary dependent treatment options, e.g. spelling variations, spaces/hyphens, diacritics, Greek letters, plural forms. This context-sensitive fine-tuning is especially important in the annotation of chemistry and protein terms.

Importantly, recognized terms are semantically interpreted as mentions of concepts that are ontologically located within domain-specific taxonomies. For chemistry named entity recognition this semantic interpretation may include the assignment of chemical structures or Markush structures to compounds or chemical class terms. OCMiner® dictionaries are generated from fine-grained domain ontologies in the form of conceptual taxonomies. This semantic mapping provides the basis for subsequent ontological indexing methods and knowledge extraction technologies.

Particular attention is given to the chemical dictionary. It is generated from a compound structure database built from various publicly available sources such as PubChem, MeSH, DrugBank, ChEMBL, among others. Our system is able to automatically arrange compounds into a single chemical ontology according to their structure or their functional properties [4]. As a consequence, a given textual expression is not only recognized as a chemical term but also semantically interpreted as a mention of a chemical entity which is precisely classified in the taxonomy. Similarly, the knowledge of other domains is hierarchically organized into taxonomies of concepts of varying specificity, e.g. proteins, genes, species, diseases or anatomy.

In parallel to the dictionary look-up of database compounds, our system also makes use of name-to-structure conversion tools (OPSIN [5], ChemAxon [6]). Before name-to-structure conversion takes place and due to widespread OCR spelling errors in patent texts, our system first tries to correct widespread spelling errors by applying textual substitutions based on regular expressions before sending potential chemical names to above name-to-structure engines.

A special module called “MolPuzzler” is dedicated to the recognition of chemical formulas. Commonly used types of chemical formulas are, among others, sum formulas (e.g. C2H5O) and condensed formulas (e.g. CH3-CH=CH2), as well as mixed forms and abbreviations of

substituent groups (e.g. “Me”) within them. Our system tries to build a valid chemical structure (e.g. SMILES [7]) from these expressions. If it succeeds, then the expression in question is very likely to be a valid chemical term.

A chemistry-specific module tries to recognize whether a given chemical expression refers to a specific compound, a compound class, or a substituent group/fragment [8]. This module considers the annotated text, information about the chemical concept it refers to, and the surrounding context. Later, the resulting knowledge on a chemical term type may be used among others for correcting annotation errors. This is especially useful in case of accumulations of various consecutive annotations which are either combined or deleted, depending on the involved chemical term types.

Additional components handle specific scenarios. For instance, the abbreviation annotator finds expansions of acronyms and abbreviated terms. Another module recognizes expressions like “vitamin A and B” as a coordinated entity and annotates “vitamin A” as such and “B” as “vitamin B”.

3 Results and Discussion

In the BioCreative V CHEMDNER-Patents challenge, our system achieved the following results [2]. In the chemical entity mention in patents (CEMP) task, precision was 0.81 at a recall of 0.76, yielding an F-score of 0.78. In the chemical passage detection (CPD) task, we obtained a sensitivity of 0.93 and a specificity of 0.88, yielding an accuracy of 0.91.

A number of conclusions can be drawn from these results. First, the annotation guidelines applied in the CHEMDNER-Patents task differ in some points from our annotation principles, especially for chemical expressions referring to more than one chemical entity. This circumstance leads to different annotations.

Second, specifically for patent texts, due to the widespread misspellings of chemical names, our system cannot assign a valid chemical structure to chemical compound expressions with spelling errors or missing parenthesis. Furthermore, patents often describe new chemical compounds which are not yet contained in our database-backed dictionary of chemical expressions.

We conclude that for the specific task in this challenge, a statistical system based on machine learning might be better suited than our rule-based, database-backed system. However, it should be noted that the task of the challenge, to recognize a text passage as a mention of a chemical entity, constitutes just the first (syntactic) step in the interpretation of textual expressions as a chemical entity. The second (semantic) step, also known as concept mapping, i.e. the actual interpretation of a recognized chemical expression as referring to a specific chemical entity to which a chemical structure can be assigned or which can be classified within a chemical ontology, has not been evaluated. It remains to be demonstrated how chemical entity recognition systems perform on the complete task of locating chemical expressions within texts and mapping them to specific chemical entities, not only recognizing a textual form but also interpreting its chemical content.

REFERENCES

1. Verberne, Suzan, Cornelis Koster and Nelleke Oostdijk (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.
2. Krallinger et al. (2015): Overview of the CHEMDNER patents task. In Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, Sevilla.
3. Apache UIMA: http://uima.apache.org/
4. Bobach, Claudia., Timo Böhme, Ulf Laube, Anett. Püschel, Lutz Weber (2012): Automated compound classification using a chemical ontology, *J. of Cheminformatics* **4**(1), 40.
5. OPSIN: http://opsin.ch.cam.ac.uk/
6. ChemAxon: http://www.chemaxon.com/
7. SMILES: http://www.daylight.com/
8. Irmer, Matthias, Claudia Bobach, Timo Böhme, Anett Püschel and Lutz Weber (2013): Using a chemical ontology for detecting and classifying chemical terms mentioned in texts. *In Proceedings of Bio-Ontologies 2013*, Berlin, http://www.bio-ontologies.org.uk/