Chemical Named Entity Recognition with OCMiner

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

OntoChem GmbH

Halle (Saale), Germany

*1<firstname>.<lastname>@ontochem.com

Abstract. We present a system for the recognition of chemical terms mentioned in texts. Our system is a modular text processing pipeline. Annotation of named entities is dictionary-based and especially suited for the rapid annotation of large document collections. The chemistry dictionary is based on a chemical ontology that is designed to be supportive of text mining and knowledge extraction tasks. The system is able to interpret chemical formulas and additionally uses name-to-structure conversion tools. Additional components help recognizing chemical terms and formulas, their type (compound classes, specific compounds and substituent groups) and their structures. Chemical named entity recognition is semantically enhanced by a variety of context-sensitive modules for annotation validation and improvement.

Keywords. Chemical named entity recognition; Chemical ontology; Chemical term type; Chemical formula recognition

1 Introduction

Chemical named entity recognition is one of the major challenges in recognizing domain-specific terms in free text. The major reason for the difficulty of correctly annotating chemical terms is the great variability of chemical expressions. There are trivial and systematic names for chemical compounds and classes, as well as trade names or international nonproprietary names (INN [1]) for drugs. Chemical names can be extremely long and may contain punctuation symbols and parentheses. There are also various types of chemical formulas or commonly known numbers such as the chemical abstracts service number (CAS [2]), or the international chemical identifier (InChI [3]) - moreover different name parts can even be mixed within one chemical expression.

We present OCMiner, a high-performance text processing system for large document collections ranging from short scientific abstracts, full-text articles and patents up to books. Several linguistic options in OCMiner allow adjusting the quality of annotation results which can be specialized and fine-tuned for the recognition of chemical terms.

2 System description

OCMiner is a modular processing pipeline for unstructured information based on the Apache UIMA framework [4]. Chemical named entity recognition (CNER) is implemented by integrating a number of different toolbox modules into the OCMiner pipeline. The general architecture is depicted in Fig. 1.

Readers process data from a variety of sources, standardizing the input for further analysis. Analysis engines add further data. Consumers provide the final output - in the case of the CHEMDNER challenge in the BioCreative annotation format.

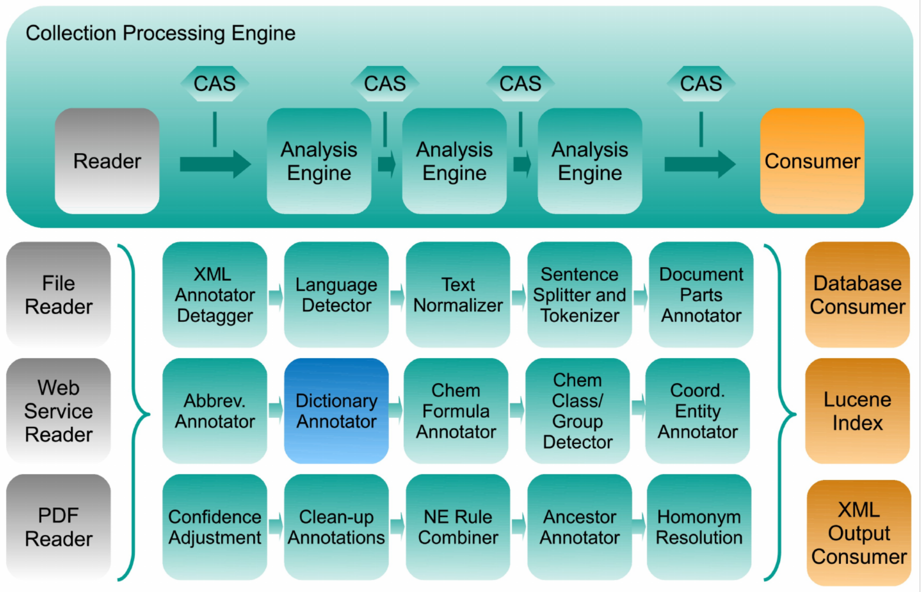

Figure 1: OCMiner UIMA pipeline for Chemical Named Entity Recognition

At the core of the system are analysis engines. First of all, there are some preparatory processes such as language detection, sentence splitting, tokenization, document structure recognition. Then, dictionary-based named entity recognition constitutes the most important module for the CHEMDNER task. We use a high performance dictionary look-up technology with support for very large dictionaries (our chemical dictionary has ca. 34 million entries) [5]. It implements specific language and dictionary dependent treatment options. For example, it is adaptable to recognize spelling variations, e.g. spaces/hyphens, diacritics, Greek letters (e.g. “α-amino acid” → “alpha-amino acid”), plural forms, apostrophe s. The dictionary can be fine-tuned with the help of domain-specific blacklists, whitelists and graylists. They include stop words, common words and homonymic expressions. In order to get meaningful annotation results, some terms have to be blacklisted in general. Other terms can be identified as chemical names in a particular context only (e.g. enumerations). These terms are collected in a "graylist" and are only annotated under specific circumstances. Furthermore, there are conditional black- and whitelists where particular context conditions can be specified.

The chemical dictionary is generated from a compound database which is built from various publicly available sources such as PubChem, MeSH, DrugBank, ChEMBL, among others. It contains chemical structures for over 14 million compounds which are stored together with their most common synonyms such as trivial and systematic names, IUPAC names, drug names. The compounds in the database are assigned to chemical compound classes using our chemistry ontology.

Our chemistry ontology constitutes the "backbone" of our system's chemical knowledge. It is an ontology of chemical classes which permits an automated high-quality hierarchical classification of chemical compounds according to their structure or their functional properties [6]. The ontology covers the most relevant upper-level chemical compound classes and substituent group classes. All of them contain chemical structure information in form of SMARTS structural patterns [7], which are the basis of a structure based automatic assignment of chemical compounds to these upper-level classes. We are thus able to annotate a chemical term referring to a compound with its ontological parent classes. This allows for an ontological search of compounds in search engines such as OCMiner [8].

Besides the dictionary look-up of database compounds, our system also makes use of name-to-structure conversion tools (OPSIN [9], ChemAxon [10]).

A special module is dedicated to the recognition of chemical formulas. Commonly used types of chemical formulas are, among others, sum formulas (e.g. C_2H_5O) and constituent formulas (e.g. CH_3-CH=CH_2), as well as mixed forms and abbreviations of substituent groups (e.g. "Me") within them. Our system tries to build a 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.

Additional components handle specific scenarios. For instance, we use an abbreviation annotator which 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". A chemistry-specific module tries to recognize whether a given chemical expression refers to a specific compound, a compound class, a substituent group, or classes thereof [11]. This module considers the annotated text, information about the chemical concept it refers to, and the surrounding context. The knowledge of a chemical term type is used 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.

3 Discussion

The annotation guidelines for the CHEMDNER task differ in certain points from the strategy used by our system. For instance, coordinated entities ("vitamin A and B") appear in the manually annotated corpus as a single entity, while our system maps them to two distint concepts. However, after minimally adapting the system to the needs of the CHEMDNER task using the training corpus, we obtained the following results on the development corpus of 3500 abstracts. For the chemical entity mention (CEM) task, we obtained a (micro-averaged) precision of 0.84 at a recall of 0.71 (F-score 0.77). For the document indexing (CDI) task, precision was 0.82 at a recall of 0.72 (F-score 0.77).

4 Acknowledgment

Part of this work was funded by a grant of the German ministry of education and research (BMBF), project "SARminer" (01IS12011A).

REFERENCES

1. INN: http://www.who.int/medicines/services/inn/en/
2. CAS: http://www.cas.org/
3. Stein, Stephen E., Stephen R. Heller, and Dmitrii Tchekhovskoi (2003): An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier, in *Proceedings of the 2003 International Chemical Information Conference (Nimes)*, Infonortics, pp. 131-143.
4. US8473501, Weber, Lutz and Timo Böhme: "Methods, Computer Systems, Software and Storage Media for Handling Many Data Elements for Search and Annotation", published March 3, 2011.
5. Apache UIMA: http://uima.apache.org/
6. Bobach, Claudia, Timo Böhme, Ulf Laube, Anett Püschel and Lutz Weber (2012): Automated compound classification using a chemical ontology, *Journal of Cheminformatics* **4**(1), 40.
7. SMILES/SMARTS: http://www.daylight.com/
8. OCMiner: http://www.ocminer.com/
9. Lowe, Daniel (2012): *Extraction of chemical structures and reactions from the literature*, PhD thesis, University of Cambridge, http://opsin.ch.cam.ac.uk/
10. Bonniot de Ruisselet, Daniel (2011): *Reliably converting names to structures with ChemAxon tools*, http://www.chemaxon.com/
11. 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/