OCMiner: Text Processing, Annotation and Relation Extraction for the Life Sciences

Timo Böhme, Matthias Irmer, Anett Püschel, Claudia Bobach,
Ulf Laube, and Lutz Weber

OntoChem GmbH, Halle (Saale), Germany
{timo.boehme,matthias.irmer,anett.pueschel,
claudia.bobach,ulf.laube,lutz.weber}@ontochem.com
http://www.ontochem.com

Abstract. We present OCMiner, a high-performance text processing system for large document collections of scientific publications. Several linguistic options allow adjusting the quality of annotation results which can be specialized and fine-tuned for the recognition of Life Science terms. Recognized terms are mapped to semantic concepts which are ontologically located within their respective domain taxonomies. Relying on a correct identification and semantic interpretation of mentions of domain concepts, relations between entities are extracted. The annotated text, as well as extracted knowledge triples, can be visualized on a web-based front-end at http://www.ocminer.com/, permitting an explorative information retrieval.

Keywords: text mining, chemical named entity recognition, relation extraction, explorative information retrieval

1 Background

Life Science knowledge mining methods rely on a correct annotation of terms and phrases with concepts from different knowledge domains – in particular chemistry, proteins and diseases – followed by the application of suitable semantic relation extraction algorithms. We present the implementation of a high quality context sensitive annotation of named entities in text documents that makes use of an exchangeable set of chemistry, protein and disease ontologies.

A particular challenge in recognizing Life Science terms in free text is chemical named entity recognition. The difficulty of correctly annotating chemical terms lies in the large number of chemical terms and chemicals as well as in the great variability of chemical expressions: There are trivial and systematic names for chemical compounds and classes, as well as formulas and trade names for drugs. Chemical names can be extremely long and may contain variations of meaningful punctuation symbols and parentheses. Moreover, different chemistry name types can even be mixed within one chemical expression. Similarly, the recognition of protein terms in texts and the correct mapping to protein concepts is a non-trivial issue. Protein terms are often abbreviated and appear

in various spelling variants (with or without hyphens, spaces etc., e.g. FLT1, FLT-1, FLT 1) and may be confused with other terms (e.g. ASK protein). Likewise, frequent disease terms are often homonymous to other concepts, e.g. a "flash" might be a physiological circumstance only in certain contexts. In sum, the precise identification of named entities is an important prerequisite for the extraction of correct and relevant relations between annotated concepts, e.g. metabolic pathways or relations between chemicals and diseases.

2 System Description

OCMiner is a modular processing pipeline for unstructured information based on the Apache UIMA framework. The system architecture is depicted in Fig. 1.

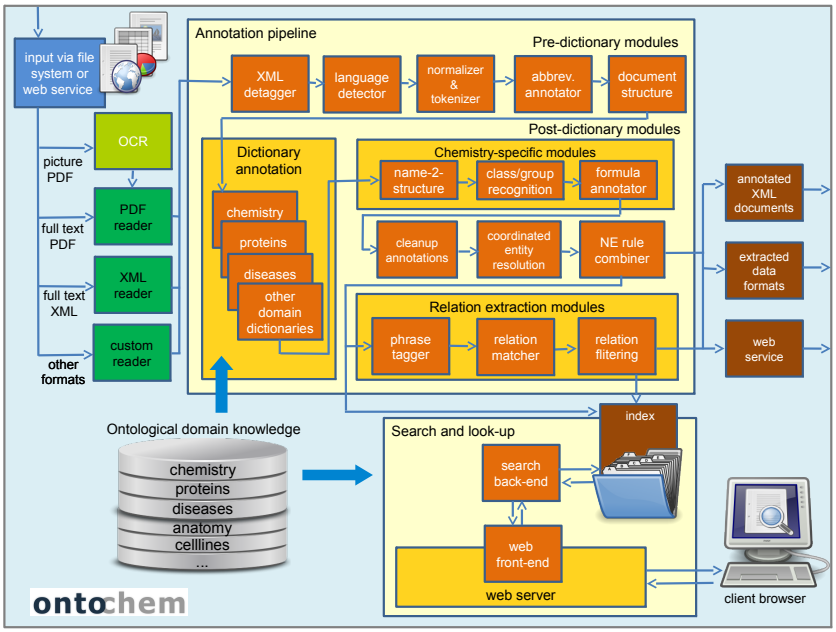

Fig. 1. OCMiner document processing pipeline

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, sentence splitting, 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 (our chemical dictionary has about 34 million entries). It 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 protein and chemistry terms.

Importantly, recognized terms are semantically interpreted as mentions of concepts that are ontologically located within domain-specific taxonomies. Our dictionaries are generated from fine-grained domain ontologies in form of conceptual taxonomies. This semantic mapping provides the basis for ontological search methods and knowledge extraction technologies. Particular importance is given to the chemical dictionary. It is generated from a compound 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 [1]. 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 specifity, eg. species, diseases or anatomy.

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". 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.

A processing step which serves as a prerequisite for relation extraction is the combination of annotated concepts to complex entities. Thus, post-dictionary modules combine sequences of named entities. For instance, the text phrase "human raf kinase inhibitor", initially annotated as a sequence of named entities [$_{organism}$ human] [$_{protein}$ raf] [$_{protein}$ kinase] [$_{chemistry}$ inhibitor], is combined to a single – though internally complex – entity referring to a chemical compound class: [$_{chemistry}$[$_{protein}$[$_{organism}$ human] raf kinase] inhibitor]. This is especially useful for the recognition of higher-level combined entities made up of constituents of the domains chemistry, proteins, species, anatomy and celllines.

Our system applies a shallow pattern-based approach to the extraction of relations between annotated concepts from different domains, e.g. between chemicals and diseases, or metabolic pathways, physico-chemical properties of compounds, etc. First, the annotated input text is tokenized into phrase tokens, where named entities, including higher-level combined entities, constitute single tokens. Note that the system does not rely on part-of-speech tagging, parsing or other sophisticated but time-consuming natural language processing techniques. Instead, extraction rules work on phrase tokens and take specific attributes of involved named entities into account, such as the type of a chemical entity. A dedicated relation ontology defines a taxonomy of relations to be extracted. Examples for relation concepts are "[compound] treats [disease]" or "[compound] metabolizes to [compound]". For each relation concept, specific mappings from natural language syntax patterns to semantic normalizations are defined. Pattern definitions have a syntax and a complexity similar to regular expressions, allowing for nested grouping and variable order of tokens. The relation matcher module matches the tokenized input text against these rules and generates nor-

malized relation representations in form of triples of concept identifiers: ⟨entity$_1$, relation concept, entity$_2$⟩.

Processed documents and extracted information can be stored in various ways. First, XML documents with inline annotations of recognized concept mentions can be generated out of heterogeneous input documents. Second, extracted knowledge such as keyword lists and relations between mentioned concepts (i.e. knowledge triples) can be stored in various formats (custom XML formats, RDF triples, SBML, CML, etc.) and accessed as a web service. Third, annotated entities and relation triples can be stored in an index (triple store or Lucene index), which is used for a web-based retrieval and visualization of the data. In particular, the index is accessed by a search back-end, which provides the indexed data to the web front-end. The OCMiner front-end displays annotated documents and provides a user interface for the navigation along relation chains, e.g. from chemicals over proteins to diseases, permitting an explorative information retrieval on multitudes of scientific publications (Fig. 2).

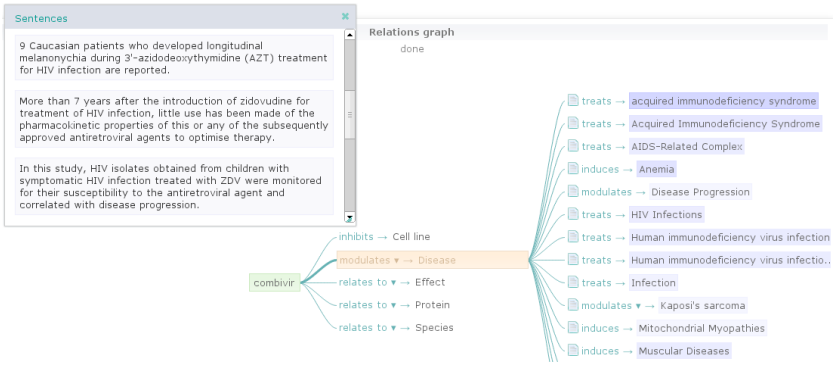

Fig. 2. OCMiner web interface for relation navigation

3 Evaluation

The system was evaluated as part of the BioCreative IV challenge. In the CHEMD-NER task for evaluating chemical NER, we obtained a precision of 85% at a recall of 71% (F-score 78%) [2]. In the CTD task, which consisted in providing annotations as a web service, our system reached an outstanding response time of 0.14 s/document, while ranking among the first two teams in annotation quality [3].

References

1. Bobach, C., T. Böhme, U. Laube, A. Püschel, L. Weber (2012): Automated compound classification using a chemical ontology, *J. of Cheminformatics* **4**(1), 40.
2. Irmer, M., C. Bobach, T. Böhme, U. Laube, A. Püschel, L. Weber (2013): Chemical Named Entity Recognition with OCMiner, *Proceedings of the 4th BioCreative challenge evaluation workshop*, vol. 2, 92-96.
3. Wiegers, T. C., A. P. Davis and C. J. Mattingly (2014): Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. *Database*. doi:10.1093/database/bau050.