Adapting the OCMiner text processing system to the CTD controlled vocabulary

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

[1]OntoChem GmbH, Halle (Saale), Germany

*Corresponding author: E-mail: lutz.weber@ontochem.com

Abstract

We adapted OCMiner, a modular text processing pipeline especially suited for high-speed processing of large document collections, to a specific controlled vocabulary as given by the Comparative Toxicogenomic Database (CTD). We provide a RESTful web service which processes documents given in the BioCreative XML format and annotates them with domain-specific terms from the CTD domains genes, chemistry, diseases and action terms.

Keywords: Text mining; Named entity recognition; Controlled vocabulary; Domain-specific annotation

Introduction

The Comparative Toxicogenomic Database (CTD) is a publicly available resource consisting of accumulated knowledge in the domains chemistry, genes/proteins, and diseases. The database contains both manually curated and automatically inferred chemical-gene/protein interactions, chemical-disease relationships, and gene-disease relationships. The BioCreative CTD task consisted in providing a web service for automatically annotating documents with domain-specific terms contained in the CTD database.

We adapted OCMiner, a modular text annotation and information extraction system especially suited for high-speed processing of large document collections, to the specific controlled vocabulary of CTD terms.

System decscription

Preparatory work

In a first step, the given CTD ontologies / taxonomies were converted into the OBO format (OBO foundry, http://www.obofoundry.org/) according to OntoChem's standard procedure. Synonyms were generated, cleaned and expanded:

- synonyms were generated from the name,
- synonyms with comma were transformed into comma-free synonyms with a term reversal around the comma: e.g. “Calculus, Kidney” into “Kidney Calculus”,
- duplicates were removed,
- OBO attributes were added to each synonym (synonym scope and type e.g. EXACT SYNONYM, as well as source, date, language and preflabel). The preflabel synonym is typically the originally supplied name.

Domain-specific blacklists, whitelists and graylists were created. They include stop words, common words and the like. Especially in the case of proteins/genes, there is a considerable amount of synonyms which are homonyms to common English words (e.g. "the", "and"). In order to get meaningful annotation results, these terms have to be blacklisted. Some terms can be identified as gene/protein synonyms 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 context conditions can be specified. For instance, the term "localization" is not annotated as an action term when preceded by "histochemical".

Processing pipeline

For the BioCreative CTD annotation task, we make use of our OCMiner text mining system. It is a high-throughput UIMA-based (http://uima.apache.org/) modular framework designed for the rapid annotation of huge collections of texts of various categories (abstracts, journal articles, patents, technical documentations) and formats (PDF, XML, HTML,…). The overall architecture is depicted in Fig. 1.

Figure 1. OCMiner UIMA pipeline

The *collection reader* reads text data form varying sources (here: XML documents via web service). Then, a number of preparatory modules make sure that the text becomes processable. First of all, XML tags are separated form the proper text. Then, whitespaces and special characters are normalized, before the text is tokenized and ready for Named Entity recognition.

The most important components in the pipeline are the dictionary-based *domain annotators*. From the converted taxonomies, a dictionary was created for each domain. The domain annotators use these dictionaries, which are designed for high-speed lookup of terms in texts. A special feature is the ability to deal with typical variations in domain term usage. For instance, the protein term "5-HT2A" may be written as "5HT2A" or "5HT-2A".

Additional components handle specific scenarios. For instance, we use an *abbreviation annotator* which finds expansions of acronyms and abbreviated terms. The *coordinated entity annotator* recognizes expressions like "vitamine A and B" as a coordinated entity and annotates "vitamine A" as such and "B" as "vitamine B".

In a *post-processing* step, annotations are validated and cleaned in a configurable manner. As a general rule, we do not allow overlapping annotations. If an annotated term (e.g. "protease") is subsumed by another annotated term (possibly from another domain, e.g. "protease inhibitor"), then only the longer term is kept. Similarly, graylisted terms become availabe as annotations if they are part of enumerations of terms from the same domain.

Finally, a *consumer* writes the annotated data in a specific format to files, databases or indexes. For the BioCreative CTD task, the original XML file is augmented with annotations at a specified position and sent as a response to the web service query.

Annotated domain	Precision (macro-averaged)	Recall (macro-averaged)	Average seconds processed
gene	0.4454	0.6727	0.1419
disease	0.4583	0.5944	0.1404
chem	0.6244	0.8540	0.1407
action term	0.3996	0.3635	0.1396

Table 1. Final challenge results for CTD annotation

Results

In the BioCreative CTD challenge, we obtained the final results given in Table 1. Note that the retrieval of action terms did not coincide very much with the manually curated information. This is due to the fact that in many cases curators tagged a document with a specific "action term" expressing a relationship chemistry/genes, genes/diseases, or chemistry/diseases, while the relationship was not explicitly expressed in the text using these terms. This suggests that a "pure" named entity recognition as applied here might not be a suitable method for relationship extraction. On the other hand, results in the other domains were much better, with best results in the chemistry domain.

Funding

This work was supported by a grant of the German ministry of education and research (BMBF), project "SARminer" [grant number 01IS12011A].