OntoGene: CTD entity and action term recognition

Fabio Rinaldi, Simon Clematide, Tilia Renate Ellendorff, Hernani Marques

Motivation and Objectives

Text mining technologies are increasingly providing an effective response to the growing demand for faster access to the vast amounts of information hidden in the literature. Several tools are becoming available which offer the capability to mine the literature for specific information, such as for example protein-protein interactions or drug-disease relationships. The biomedical text mining community regularly verifies the progress of such systems through competitive evaluations, such as BioCreative, BioNLP, i2b2, CALBC, CLEF-ER, BioASQ, etc.

The OntoGene system is a text mining system which specializes in the detection of entities and relationships from selected categories, such as proteins, genes, drugs, diseases, chemicals. The quality of the system has been verified several times through participation in some of the combg munity-organized evaluation campaigns, where OntoGene has consistently achieved top-ranked results. Some highlights include best results in the detection of experimental methods (BioCreative 2006), best results in the detection of protein-protein interactions (BioCreative 2009), best results in large-scale detection of some categories of biomedical entities (CALBC 2010), best overall results in the CTD triage task (BioCreative 2012).

However, the OntoGene system is based on a relatively heterogeneous pipeline, which would not be easily portable to other sites. In order to make the advanced text mining capabilities of the OntoGene system more widely accessible to the biomedical community without the burden of installation of complex software, we long planned to provide access through web services.

The task 3 of BioCreative 2013 provided the ideal setting to implement an initial version of such web service interface. The goal of task 3 was to deliver entity and action term annotation for the Comparative Toxicogenomics Database (3).

Methods

The text mining pipeline which constitutes the core of the OntoGene system has been described previously in a number of publications (5,6). We will only briefly describe the core text mining technologies, and instead focus mainly on the novel web service which allows remote access to the OntoGene text mining capabilities.

The first step in order to process a collection of biomedical literature consists in the annotation of names of relevant domain entities in biomedical literature (currently the systems considers

proteins, genes, species, experimental methods, cell lines, chemicals, drugs and diseases). These names are sourced from reference databases and are associated with their unique identifiers in those databases, thus allowing resolution of synonyms and cross-linking among different resources. A term normalization step is used to match the terms with their actual representation in the text, taking into account a number of possible surface variations. Finally, a disambiguation step resolves the ambiguity of the matched terms.

A supervised machine learning method is used to generate a score for entity annotations. Since the term recognizer aims at high recall, it introduces several noisy concepts, which we want to automatically identify in order to penalize them. Additionally, we need to adapt to highly-ranked false positive relations which are generated by our frequency based approach. The goal is to identify some global preference or biases which can be found in the reference database. Our technique is to weight individual concepts according to their likeliness to appear as an entity in a correct relation, as seen in the target database. The same approach was previously used for our participation in BioCreative 2012 (8). The only adaptation was to use the most recent version of the CTD datasets for training (about 97'000 pubmed articles), filtered by the criterion that there were not more than 12 relations curated in an article. This led to a number of 328230 curated relations from these articles where we applied the supervised distant learning approach for scoring the concept relevance. The term database for genes, chemicals and diseases has 454,429 concepts and 1,282,582 terms.

The OntoGene web service has been implemented as a RESTful service (2). It accepts simple XML files as input, based on the BioC specification[1]. The output of the system is generated in the same format. Options can be used in the input query to select whether the result should contain in-line annotations (showing where exactly in the text the term was mentioned), or stand-off annotations. Currently the system uses pre-defined terminology, however we foresee in future the possibility to upload own terminologies, or select which subsets of the available terminology should be used.

Action Terms

In order to be able to discover action terms in unseen abstract, we built several binary machine-learning classifiers, one for each action-term type. We did not use the ontogene pipeline for building the classifiers, but decided to base our system mainly on tools from the natural language processing toolkit NLTK (1). As training material, we made use of the official CTD data for gene-chemical interaction which can be downloaded from the website in xml-format as well as the referenced abstracts from pubmed. In addition to the abstracts, we used the MeSH descriptors and qualifiers as PubMed metadata. Any preprocessing of the abstracts was not done, apart from sentence splitting and tokenization.

[1] http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/BioC/

For building classifiers which were able to classify abstracts according to the action terms which they contain, we used the Naive Bayes classifier module, provided by NLTK. For each action term to be classified, one binary classifier was built. All features used are independent binary features, as required by the Naive Bayes theorem. We programmed the classifiers in a way that made it possible to use different combinations of feature-types. The simplest feature-type only uses the words of the abstract (bag-of-words). Other feature-types considered the stems of the words, word-bigrams, stem-bigrams, mesh-descriptors and mesh-qualifiers. Furthermore, the number of most frequent features used for a feature-type could be adjusted. Experiments showed that using the 5000 most frequent features for each feature type (e.g. the 5000 most frequent words are used as features) leads to the best results.

Another setting that we varied in order to find optimal performance is the number of action term types for which classifiers were included. Out of a total of 53 action terms, we made experiments with systems including classifiers for from 7 to 15 action terms. The best performance in terms of F-Score could be measured for the system which included 9 different classifiers.

The last variable of the classifying system was the size of its training set which consisted of abstracts randomly chosen from the total number of pubmed abstracts listed in CTD. Here it is important to take the efficiency of the system into account: the classifier tends to run very slow if too much data is provided, without big improvement once a certain amount of data is reached. We found that a training set of 2000 different abstracts shows a reasonably good performance together with an adequate speed rate.

With the help of experiments using different feature settings, we determined the best choice of features as bag-of-words, stem bigrams and mesh descriptors. In this context we found that mesh descriptors are the most useful features for determining action words followed by stem-bigrams. (Even though word-bigrams were found to be still a bit more useful than bag-of-words, using both at the same time seems to introduces redundant information and leads to a worse performance, the same seems to be the case with using stems together with stem-bigrams.) Using Mesh-qualifiers together with Mesh-descriptors as one feature proved to be too sparse to have any positive effect.

Results and Discussion

Users can submit arbitrary documents to the OntoGene mining service by embedding the text to be mined within a simple XML wrapper. Both input and output of the system are defined according to the BioC standard (2). However typical usages will involve processing of PubMed abstracts or PubMed Central full papers. In this case the user can provide as input simply the pubmed identifier of the article. Optionally the users can specify which type of output they would like to obtain: if entities, which entity types, and if relationships, which combination of types.

The OntoGene pipeline identifies all relevant entities mentioned in the paper, and their interactions, and reports them back to the user as a ranked list, where the ranking criteria is the system own confidence in the specific result. The confidence value is computed taking into account several factors, including the relative frequency of the term in the article, its general frequency in PubMed, the context in which the term is mentioned, and the syntactic configuration among two interacting entities (for relationships). A detailed description of the factors that contribute to the computation of the confidence score can be found in (6).

The user can chose to either inspect the results, using the ODIN web interface (see figure 1), or to have them delivered back via the RESTful web service in BioC XML format, for further processing locally. The usage of ODIN as a curation tool has been tested within the scope of collaborations with curation groups, e.g. PharmGKB (7).

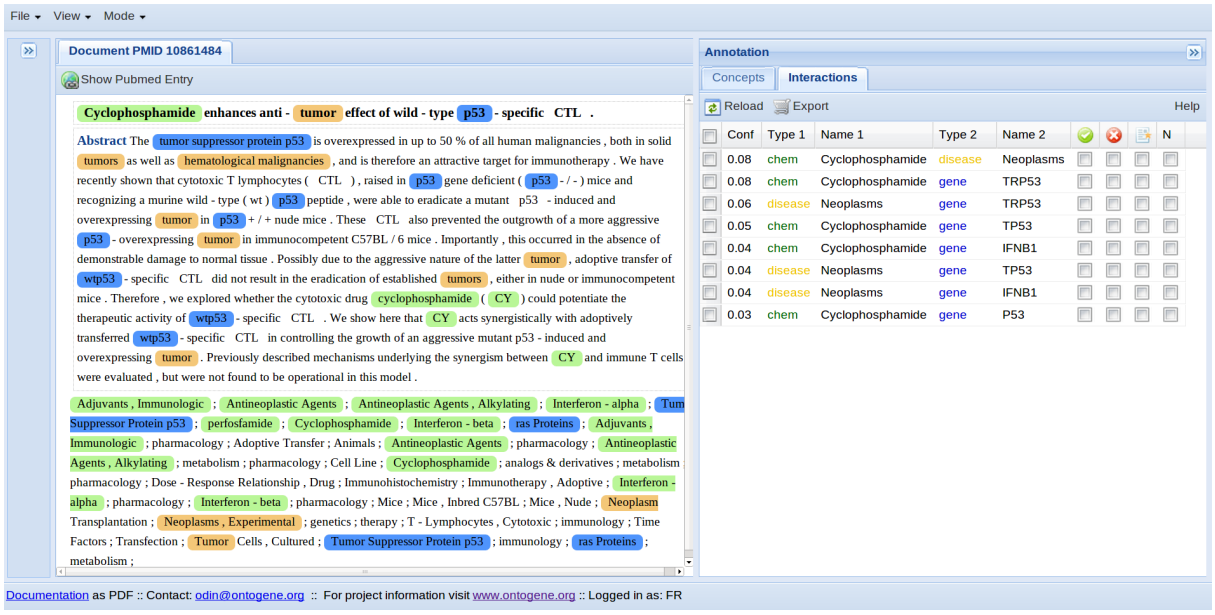

Figure 1. Example of visualization of text mining results using the ODIN interface.

Acknowledgements

The OntoGene group is partially supported by the Swiss National Science Foundation (grants 100014– 118396 / 1 and 105315– 130558 / 1). A continuation of this work is planned within the scope of a collaboration with Roche Pharmaceuticals.

References

1. Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python. O'Reilly Media Inc.
2. Donald C. Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifang Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, W. John Wilbur (2013). BioC: a minimalist approach to interoperability for biomedical text processing, The Journal of Biological Databases and Curation (2013), *bat064, doi:10.1093/database/bat064, published online.*
3. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: update 2013. Nucleic Acids Res. 2013 Jan 1;41(D1):D1104-14.
4. Richardson, Leonard; Ruby, Sam (2007), *RESTful Web Services*, O'Reilly, ISBN 978-0-596-52926-0
5. Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, Therese Vachon (2008). OntoGene in BioCreative II. Genome Biology, 2008, 9:S13, PMC2559984
6. Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Thérèse Vachon, Martin Romacker, "OntoGene in BioCreative II.5," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(3), pp. 472-480, 2010. doi:10.1109/TCBB.2010.50
7. Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, Russ B. Altman. *Using ODIN for a PharmGKB revalidation experiment. The Journal of Biological Databases and Curation*, Oxford Journals, 2012, bas021; doi:10.1093/database/bas021
8. Fabio Rinaldi and Simon Clematide and Simon Hafner and Gerold Schneider and Gintare Grigonyte and Martin Romacker and Therese Vachon. Using the OntoGene pipeline for the triage task of BioCreative 2012. The Journal of Biological Databases and Curation, Oxford Journals, 2013. doi:10.1093/database/bas053