The OntoGene system: an advanced information extraction application for biological literature

Rinaldi, Fabio

The ontogene system: an advanced information extraction application for biological literature

Fabio Rinaldi✉
Institute of Computational Linguistics, University of Zurich

Motivation and Objectives

The rapid expansion of the biomedical knowledge encoded in the scientific literature is proving to be a major bottleneck for the progress of biomedical sciences. It is increasing difficult even for the best experts to keep track of all relevant information pertinent to their domain of interest. It is becoming therefore imperative to explore solutions based on advanced text mining technologies in order to identify and extract the most relevant nuggets of information from the vastness of the literature.

Methods

The OntoGene system (http://www.ontogene.org/) is an advanced NLP-based pipeline capable of efficiently processing large quantity of textual documentation and extracting from it specific items of information, and in particular the biomedical entities of interest to the user, and their relationships.

Biomedical terminological resources can be leveraged for construction of large-scale knowledge bases. One example is KaBOB (Knowledge Base of Biology), a large RDF store based upon 17 prominent biomedical databases (Bada et al, 2011). Similar kinds of integrated data networks can be used for knowledge discovery purposes through usage of semantic web technologies (Chen et al, 2009). In our own work we have used such databases as knowledge sources for the process of semi-automated information extraction. In the rest of this section we describe the OntoGene Text Mining pipeline which is used to (a) provide all basic preprocessing (e.g. tokenization) of the target documents, (b) identify all mentions of domain entities and normalize them to database identifiers, and (c) extract candidate interactions.

We use in particular the following resources as terminology sources: UniProt Knowledge base (proteins), NCBI Taxonomy (species), Proteomics Standards Initiative Molecular Interactions Ontology (experimental methods), Cell Line Knowledge Base (cell lines), UMLS (diseases), etc. Terms, i.e. preferred names and synonyms, are automatically extracted from the original database and stored in a common internal format, together with their unique identifiers (as obtained from the original resource). An efficient lookup procedure is used to annotate any mention of a term in the documents with the ID(s) to which it corresponds. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the list of known terms at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings (Rinaldi et al, 2008).

The system combines mentions of relevant domain entities (and their corresponding unique identifiers) from the same syntactic context in order to create candidate interactions. An initial ranking of the candidate relations can be generated on the basis of frequency of occurrence of the respective entities only. This ranking is further refined using a syntax-based approach, which is based upon an accurate parsing of all the sentences of the target document, and a machine learning approach which makes use of a maximum entropy classifier to boost candidate entities and interactions on the basis of the global distribution of information in the original database (Rinaldi, Schneider, *et al*, 2012).

Results and Discussion

The results of the text mining system are presented to the user through an intuitive and user-friendly interface, called ODIN (OntoGene Document Inspector). The ODIN interface allows the user to inspect entities and relationships identified by the text mining system, and see them in the context where they were originally found.

For example, the figure below shows an implementation of ODIN customized for curation of the Comparative Toxicogenomics Database (CTD, Mattingly et al, 2006). The left panel shows

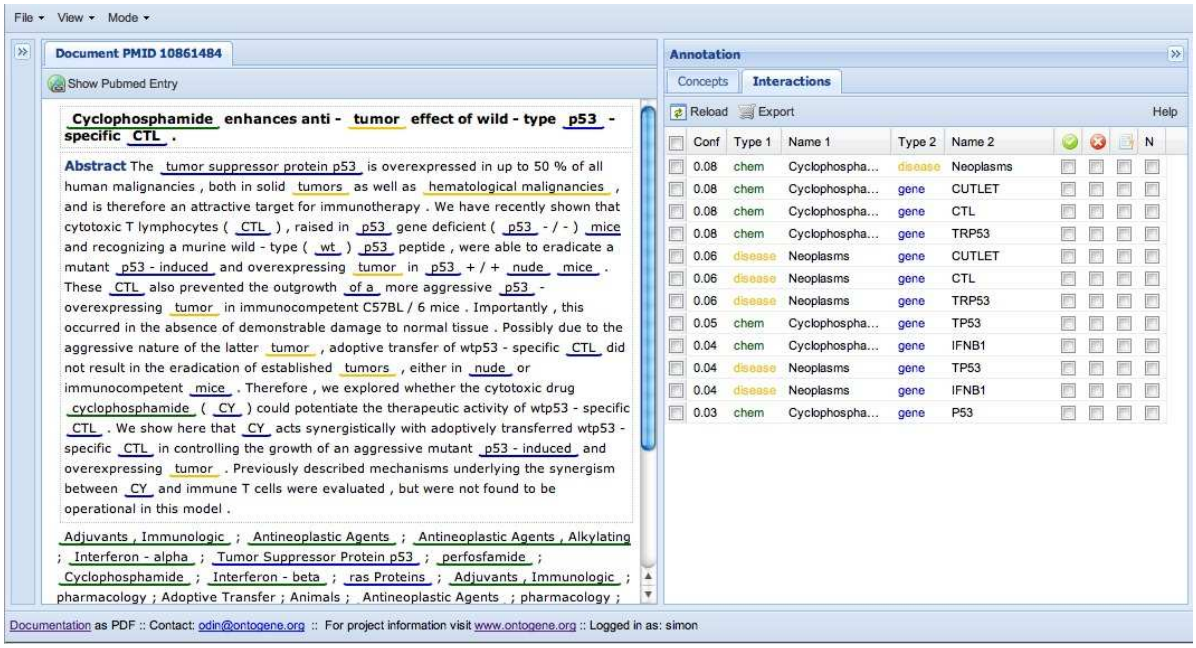

Figure 1: ODIN: the OntoGene curation interface in the CTD application.

the original document, with entities underlined and color-coded (green: chemicals, yellow: diseases, blue: genes). The right panel shows the candidate relationships identified by the system. Selecting one of the interactions will highlight in the document the information that was used by the system to propose that interaction.

The results (interactions in this case) are presented according to a ranking which is based upon a score reflecting the confidence of the system in a given proposed interaction, thus allowing the user to stop inspecting them at an optional confidence threshold. The user can with a simple click then confirm or reject a candidate interaction. Additionally, all entities are easily editable, allowing correction of annotation errors.

The OntoGene pipeline has been applied to several Information Extraction tasks. In the context of the BioCreative challenges (Krallinger et al 2008), the system was capable of achieving the best results in extracting mentions of protein-protein interactions (2009) and mentions of experimental methods for protein interaction detection (2006).

Recently the system has been adapted for an experiment in assisted curation for the PharmGKB database (Klein et al 2001). This experiment, conducted in collaboration with PharmGKB curators, has lead to interesting results showing the reliability and usability of the system (Rinaldi, Clematide, et al, 2012).

In the "triage" task of BioCreative 2012 (ranking of documents according to relevance for the curation process of the CTD database), once again the OntoGene system obtained the best overall results among the participants (Rinaldi et al, 2013).

Acknowledgements

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1) and Novartis AG, NIBR-IT, TextMining Services, CH-4002, Basel, Switzerland.

References

1. Michael Bada, Kevin Livingston, and Lawrence Hunter. An ontological representation of biomedical data sources and records. Bio-Ontologies, 2011.
2. Huajun Chen, Li Ding, Zhaohui Wu, Tong Yu, Lavanya Dhanapalan, and Jake Y. Chen. Semantic web for integrated network analysis in biomedicine. Briefings in Bioinformatics, 10(2):177–192, 2009
3. T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, and R.B. Altman. Integrating genotype and phenotype information: An overview of the PharmGKB project. The Pharmacogenomics Journal, 1:167–170, 2001.
4. M. Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, Alfonso Valencia. Evaluation of text-mining systems for biology: overview of the second BioCreative community challenge. Genome Biology, 2008, 9(Suppl 2):S1.

5. C.J. Mattingly, M.C. Rosenstein, G.T. Colby, J.N. Forrest Jr, and J.L. Boyer. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. Journal of Experimental Zoology Part A: Comparative Experimental Biology, 305A(9):689–692, 2006.
6. Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. Genome Biology, 9(Suppl 2):S13, 2008.
7. Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. Using ODIN for a PharmGKB revalidation experiment. The Journal of Biological Databases and Curation, Oxford Journals, 2012
8. Fabio Rinaldi, Gerold Schneider, Simon Clematide. Relation Mining Experiments in the Pharmacogenomics Domain. Journal of Biomedical Informatics, 2012.
9. Fabio Rinaldi, Simon Clematide, Simon Hafner, Gerold Schneider, Gintare Grigonyte, Martin Romacker, Therese Vachon. Ranking of CTD articles and interactions using the OntoGene pipeline. The Journal of Biological Databases and Curation, Oxford Journals, 2013 (accepted for publication).
10. Thomas C. Wiegers, Allan Peter Davis, and Carolyn J. Mattingly. Collaborative Biocuration-Text Mining Development Task for Document Prioritization for Curation. The Journal of Biological Databases and Curation, Oxford Journals, 2013 (accepted for publication).