Mining complex Drug/Gene/Disease relations

Rinaldi, Fabio ; Schneider, G ; Clematide, S

Mining complex Drug/Gene/Disease relations.

Fabio Rinaldi, Gerold Schneider, Simon Clematide

University of Zurich, Switzerland

We describe a recent development of the OntoGene Relation Miner system (OG-RM), aimed at the automatic extraction of drug/gene/diseases relationships from the biomedical literature. The OG-RM system was developed originally for the extraction of protein-protein interactions (PPI) from the biomedical literature [1]. The system has been tested in a controlled setting by participation to the PPI tasks of the BioCreative II and BioCreative II.5 competitive evaluations of text mining systems [2,3]. Recently, within the context of the SASEBio project (a collaboration between the OntoGene group at the University of Zurich and the NITAS/TMS group of Novartis Pharma AG), the OntoGene Relation Miner has been extended in order to deal with larger classes of entities and relationships. In particular, terminology for drugs and diseases has been derived from the Pharmacogenomics Knowledge Base (PharmGKB) [4].

The OntoGene system takes as input a document, provided either in plain text or in a number of supported xml formats (e.g. PubMed Central), and, after conversion to an internal xml-based format, applies a number of processing steps, enriching it with different categories of annotations. The processing steps include standard linguistic preprocessing (sentence boundary detection, tokenization, lemmatization, part-of-speech tagging), named entity recognition, phrase chunking, and syntactic parsing (using our own dependency-based parser) [5]. Entities detected in the input document are disambiguated with respect to a reference database (UniProt for proteins, EntrezGene for genes, NCBI taxonomy for species, PSI-MI ontology for experimental methods, PharmGKB for drugs and diseases). Drugs and disease names present a lower ambiguity degree compared to proteins and genes, however some problematic cases remain (e.g. "Leukemia" can be used as a term for 32 distinct diseases, according to PharmGKB).

Candidate interactions are generated on the basis of co-occurrence in a specified text unit. Validation of candidate interactions is based on a combination of local and global information. In particular, the presence of some salient clue words in the local context, and the syntactic structure of the fragment of the sentence connecting the candidate interactants, are used as indicators of relevance. Global information, such as frequency of mentions of the interactants in the paper, is further used for weighting purposes. As in the case of the BioCreative setup, not all relationships mentioned in a paper are relevant, but solely those that are reported by the authors as their main research results. The identification of such 'curatable' interactions is based on a classifier which attempts to distinguish sentences that report background information from sentences that describe novel results. While PharmGKB would allow construction of a test corpus of 26122 drug/disease/gene relationship over a set of 5062 distinct PubMed articles, in practice our experiments are limited by the absence of a common full text format. Currently we are evaluating the effectiveness of our approach using a manually selected subset of those articles. Detailed results will be presented at the workshop.

Although it is impossible to automatically reproduce the relationship annotations provided by PharmGKB, we think that advanced text mining techniques will soon be accepted as an important support tool for the curation process. To this aim, we have also developed a customizable interface (ODIN: OntoGene Document Inspector) which allows an effective interaction of the human curator with the underlying text mining system. A presentation/demo will be provided at the workshop.

References

[1] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, Martin Romacker. An environment for relation mining over richly annotated corpora: the case of GENIA. BMC Bioinformatics 2006, 7(Suppl 3):S3.
[2] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, Therese Vachon. OntoGene in BioCreative II. Genome Biology, 2008, 9:S13.
[3] 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.
[4] 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 (2001) 1, 167-170.
[5] Gerold Schneider, 2008. Hybrid Long-Distance Functional Dependency Parsing. Doctoral Thesis. Institute of Computational Linguistics, University of Zurich.