Semi-Automated Semantic Annotation of the Biomedical Literature

Fabio Rinaldi

Institute of Computational Linguistics, University of Zurich
fabio.rinaldi@uzh.ch

Abstract. Semantic annotations are a core enabler for efficient retrieval of relevant information in the life sciences as well in other disciplines. The biomedical literature is a major source of knowledge, which however is underutilized due to the lack of rich annotations that would allow automated knowledge discovery.
We briefly describe the results of the SASEBio project (Semi Automated Semantic Enrichment of the Biomedical Literature) which aims at adding semantic annotations to PubMed abstracts, in order to present a richer view of the existing literature.

1 Introduction

The scientific literature contains a wealth of knowledge which however cannot be easily used automatically due to its unstructured nature. In the life sciences, the problem is so acutely felt that large budgets are invested into the process of literature curation, which aims at the construction of structured databases using information mostly manually extracted from the literature. There are several dozens of life science databases, each specializing on a particular subdomain of biology. Examples of well-known biomedical databases are UniProt (proteins), EntrezGene (genes), NCBI Taxonomy (species), IntAct (protein interactions), BioGrid (protein and genetic interactions), PharmGKB (drug-gene-disease relations), CTD (chemical-gene-disease relations), and RegulonDB (regulatory interactions in E. coli).

The OntoGene group[1] aims at developing text mining technologies to support the process of literature curation, and promote a move towards *assisted curation*. By assisted curation we mean a combination of text mining approaches and the work of an expert curator, aimed at leveraging the power of text mining systems, while retaining the high quality associated with human expertise. We believe that it is possible to gradually automate much of the most repetitive activities of the curation process, and therefore free up the creative resources of the curators for more challenging tasks, in order to enable a much more efficient and comprehensive curation process. Our text mining system specializes in the detection of entities and relationships from selected categories, such as proteins, genes, drugs, diseases, chemicals. OntoGene derives some of its resources from life sciences databases, thus allowing a deeper connection between the unstructured information contained in the literature and the structured information contained in databases. The quality of the system has been tested several times through participation in some of the community-organized evaluation campaigns, where it often obtained top-ranked results. We have also implemented a platform for assisted curation called ODIN (OntoGene Document INspector) which aims at serving the needs of the curation community. The usage of ODIN as a tool for assisted curation has been tested within the scope of collaborations with curation groups, including PharmGKB [7], CTD [8], RegulonDB [5].

Assisted curation is also of utility in the process of pharmaceutical drug discovery. Many text mining tasks in drug discovery require both high precision and high recall, due to the importance of comprehensiveness and quality of the output. Text mining algorithms, however, cannot often achieve both high precision and high recall, sacrificing one for the other. Assisted curation can be paired with text mining algorithms which have high recall and moderate precision to produce results that are amenable to answer pharmaceutical problems with only a reasonable effort being allocated to curation.

[1] http:/www.ontogene.org/

Methods

The Ontogene system is based on a pipeline architecture (see figure 1), which includes, among others, modules for entity recognition and relation extraction. Some of the modules are rule-based (e.g. lexical lookup with variants) while others use machine-learning approaches (e.g. maximum entropy techniques). The initial step consists in the annotation of names of relevant domain entities in biomedical literature (currently the system 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.

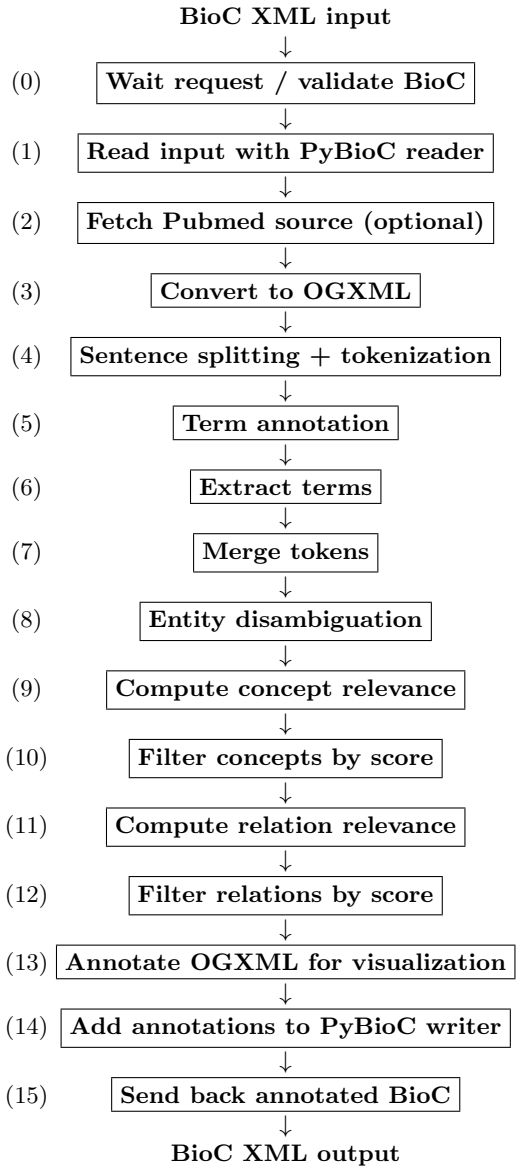

Fig. 1. Schema of the OntoGene pipeline

One of the problems with sourcing resources from several databases is the possible inconsistencies among them. The fact that domain knowledge is scattered across dozens of data sources, occasionally also with some incompatibilities among them, is a severe problem in the life sciences. Ideally these resources should be integrated in a single repository, as some projects are attempting to do (e.g. OpenPhacts [16]), allowing querying within an unified platform. However, a deep integration of the information provided by the scientific literature and the content of the databases is still missing.

We train our system using the knowledge provided by life sciences databases as our gold standard, instead of hand-labeled corpora, since we believe that the scope and size of manually annotated corpora, however much effort has been invested in creating them, is not sufficient to capture the wide variety of linguistic phenomena that can be encountered in the full corpus of biomedical literature, let alone other types of documents, such as internal scientific reports in the pharma industry, which are not represented at all in annotated corpora. For example, PubMed currently contains more than 23 million records, while the entire set of all annotated publications probably barely reaches a few thousands, most of them sparsely annotated for very specific purposes.

We generate interaction candidates using co-occurence of entities within selected syntactic units (typically sentences). An additional step of syntactic parsing using a state-of-the-art dependency parser allows us to derive specialized features in order to increase precision. The details of the algorithm are presented in [14]. The information delivered by the syntactic analysis is used as a factor in order to score and filter candidate interactions based on the syntactic fragment which connects the two participating entities. All available lexical and syntactic information is used in order to provide an optimized ranking for candidate interactions. The ranking of relation candidates is further optimized by a supervised machine learning method described in detail in [2].

Results

The OntoGene annotator offers an open architecture allowing for a considerable level of customization so that it is possible to plug in in-house terminologies. We additional provide access to some of our text mining services through a RESTful interface.[2] 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 [4]. However, typical usage 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 user 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's own confidence for 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 [14].

The user can choose to either inspect the results, using the ODIN web interface, or to have them delivered back via the RESTful web service in BioC XML format, for further local processing. ODIN (OntoGene Document Inspector) is a flexible browser-based client application which interfaces with the OntoGene server. The curator can use the features provided by ODIN to visualize selected annotations, together with the statements from which they were derived, and, if necessary, add, remove or modify them. Once the curator has validated a set of candidate annotations, they can be exported, using a standard format (e.g. CSV, RDF), for further processing by other tools, or for inclusion in a reference database, after a suitable format conversion. In case of ambiguity, the curator is offered the opportunity to correct the choices made by the system, at any of the different levels of processing: entity identification and disambiguation, organism selection, interaction candidates. The curator can access all the possible readings given by the system and select the most accurate.

As a way to verify the quality of the core text mining functionalities of the OntoGene system, we have participated in a number of text mining evaluation campaigns [9, 3, 12, 13]. Some of most interesting results include best results in the detection of protein-protein interactions in BioCreative 2009 [14], top-ranked results in several tasks of BioCreative 2010 [15], best results in the triage task of BioCreative 2012 [9]. The usage of ODIN as a curation tool has been tested in a few collaborations with curation groups, including PharmGKB [10], CTD [7], RegulonDB [11]. Assisted curation is also one of the topics being evaluated at the BioCreative competitions [1], where OntoGene/ODIN participated with favorable results. The effectiveness of the web service has been recently evaluated within the scope of one of the BioCreative 2013 shared tasks [6]. Different implementations can rapidly be produced upon request.

Since internally the original database identifiers are used to represent the entities and interactions detected by the system, the annotations can be easily converted into a semantic web format, by using a reference URI for each domain entity, and using RDF statements to express interactions. While it is possible to access the automatically generated annotations for further processing by a reasoner or integrator tool, we strongly believe that at present a process of semi-automated validation is preferable and would lead to better data consistency.

Acknowledgments. The OntoGene group is partially supported by the Swiss National Science Foundation (grant 105315 − 130558/1 to Fabio Rinaldi) and by the Data Science Group at Hoffmann-La Roche, Basel, Switzerland.

[2] http://www.ontogene.org/webservices/

References

1. Arighi, C., Roberts, P., Agarwal, S., Bhattacharya, S., Cesareni, G., Chatr-aryamontri, A., Clematide, S., Gaudet, P., Giglio, M., Harrow, I., Huala, E., Krallinger, M., Leser, U., Li, D., Liu, F., Lu, Z., Maltais, L., Okazaki, N., Perfetto, L., Rinaldi, F., Saetre, R., Salgado, D., Srinivasan, P., Thomas, P., Toldo, L., Hirschman, L., Wu, C.: Biocreative iii interactive task: an overview. BMC Bioinformatics 12(Suppl 8), S4 (2011), http://www.biomedcentral.com/1471-2105/12/S8/S4
2. Clematide, S., Rinaldi, F.: Ranking relations between diseases, drugs and genes for a curation task. Journal of Biomedical Semantics 3(Suppl 3), S5 (2012), http://www.jbiomedsem.com/content/3/S3/S5
3. Clematide, S., Rinaldi, F., Schneider, G.: Ontogene at calbc ii and some thoughts on the need of document-wide harmonization. In: Proceedings of the CALBC II workshop, EBI, Cambridge, UK, 16-18 March (2011)
4. Comeau, D.C., Doan, R.I., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T.C., Wu, C.H., Wilbur, W.J.: BIoC: a minimalist approach to interoperability for biomedical text processing. The Journal of Biological Databases and Curation bat064 (2013), published online
5. Gama-Castro, S., Rinaldi, F., Lpez-Fuentes, A., Balderas-Martnez, Y.I., Clematide, S., Ellendorff, T.R., Collado-Vides, J.: Assisted curation of growth conditions that affect gene expression in e. coli k-12. In: Proceedings of the Fourth BioCreative Challenge Evaluation Workshop. vol. 1, pp. 214–218 (2013)
6. Rinaldi, F., Clematide, S., Ellendorff, T.R., Marques, H.: OntoGene: CTD entity and action term recognition. In: Proceedings of the Fourth BioCreative Challenge Evaluation Workshop. vol. 1, pp. 90–94 (2013)
7. Rinaldi, F., Clematide, S., Garten, Y., Whirl-Carrillo, M., Gong, L., Hebert, J.M., Sangkuhl, K., Thorn, C.F., Klein, T.E., Altman, R.B.: Using ODIN for a PharmGKB re-validation experiment. Database: The Journal of Biological Databases and Curation (2012)
8. Rinaldi, F., Clematide, S., Hafner, S.: Ranking of ctd articles and interactions using the ontogene pipeline. In: Proceedings of the 2012 BioCreative workshop. Washington D.C. (April 2012)
9. Rinaldi, F., Clematide, S., Hafner, S., Schneider, G., Grigonyte, G., Romacker, M., Vachon, T.: Using the OntoGene pipeline for the triage task of BioCreative 2012. The Journal of Biological Databases and Curation, Oxford Journals (2013)
10. Rinaldi, F., Clematide, S., Schneider, G., Romacker, M., Vachon, T.: ODIN: An advanced interface for the curation of biomedical literature. In: Biocuration 2010, the Conference of the International Society for Biocuration and the 4th International Biocuration Conference. p. 61 (2010), available from Nature Precedings http://dx.doi.org/10.1038/npre.2010.5169.1
11. Rinaldi, F., Gama-Castro, S., Lpez-Fuentes, A., Balderas-Martnez, Y., Collado-Vides, J.: Digital curation experiments for regulondb. In: BioCuration 2013, April 10th, Cambridge, UK (2013)
12. Rinaldi, F., Kappeler, T., Kaljurand, K., Schneider, G., Klenner, M., Clematide, S., Hess, M., von Allmen, J.M., Parisot, P., Romacker, M., Vachon, T.: OntoGene in BioCreative II. Genome Biology 9(Suppl 2), S13 (2008), http://genomebiology.com/2008/9/S2/S13
13. Rinaldi, F., Kappeler, T., Kaljurand, K., Schneider, G., Klenner, M., Hess, M., von Allmen, J.M., Romacker, M., Vachon, T.: OntoGene in Biocreative II. In: Proceedings of the II Biocreative Workshop (2007)
14. Rinaldi, F., Schneider, G., Kaljurand, K., Clematide, S., Vachon, T., Romacker, M.: OntoGene in BioCreative II.5. IEEE/ACM Transactions on Computational Biology and Bioinformatics 7(3), 472–480 (2010)
15. Schneider, G., Clematide, S., Rinaldi, F.: Detection of interaction articles and experimental methods in biomedical literature. BMC Bioinformatics 12(Suppl 8), S13 (2011), http://www.biomedcentral.com/1471-2105/12/S8/S13
16. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open phacts: semantic interoperability for drug discovery. Drug Discovery Today 17(2122), 1188 – 1198 (2012), http://www.sciencedirect.com/science/article/pii/S1359644612001936