Combining Syntactic and Ontological Knowledge to Extract Biologically Relevant Relations from Scientific Papers

Anália Lourenço[1], Hugo Costa[1,2], Sónia Carneiro[1], Rafael Carreira[1,2], Miguel Rocha[2], Eugénio Ferreira[1], Isabel Rocha[1]

[1] IBB - Institute for Biotechnology and Bioengineering, Centre of Biological Engineering
[2] Department of Informatics / CCTC
University of Minho
Campus de Gualtar, 4710-057 Braga – PORTUGAL

{analia,soniacarneiro,rafaelcc,ecferreira,irocha}@deb.uminho.pt,
hugocosta.bio@gmail.com,mrocha@di.uminho.pt

Abstract

Bringing biologists and text miners closer together is a major aim towards the general usage of literature mining tools. Our contribution to this aim is an end-user tool for the extraction of problem-specific biologically relevant relations. Development efforts are being focused on easy-to-use text mining workflows including commonly available entity recognisers and syntactic processors, and the construction of a user-friendly environment that enables problem-specific tailoring by biologists.

1 Introduction

The increasing body of scientific text and the complex analysis requirements brought by systems-level approaches (namely, the integration of literature and high-throughput data) urge for automated literature curation processes [1, 2]. However, for the common biologist, literature mining is still out of reach. At most, he/she associates the concept to keyword-based searches in PubMed.
Close collaboration between biologists and text miners needs therefore to be encouraged. The deployment of most text mining tools requires programming and/or tuning (e.g. parameter selection) that are too specific for non-developers. In turn, automatic text processing must be able to deal with different biological problems and outputs have to be readily understandable to biologists. Biologist guidelines are required in key text mining tasks such as the assessment of document relevance, the selection of entities to be annotated and the evaluation of extracted relations. Moreover, manual curation of the outputs ensures the quality of the extracted information and may even help to refine some annotation processes (e.g. the recognition of previously unknown entities).

Our development efforts address this gap by delivering an end-user environment that brings together the skills of current entity recognisers and syntactic processors towards relation extraction. Through an intuitive graphical interface the biologist is able to interact with NLP tools and to perform domain-specific ontological contextualisation. Loaders for several entity annotation schemas (e.g. GENIA [3, 4], BioInfer [5], AIMed [6], Yapex [7] and @Note [8]) and the Gene Ontology (GO) are provided. Natural Language Processing (NLP) capabilities are granted by the GATE language engineering software [9] and a common relation extraction workflow is already set.
Biologists are able to tailor the text mining process to problem-specific contexts by filtering irrelevant relations and analysing different relation properties. Ontology-based semantic mapping enables biologists to incorporate their domain expertise, contextualising/customising the analysis of general outputs. Outputs on relation frequency and entity (class) co-occurrence may help to characterise important biological events. Also, he/she may look into biological participants that trigger or are affected by particular events. Relations are linked to original text passages and additional queries to external databases are also possible.
The software is open-source and is freely available at http://sysbio.di.uminho.pt/anote/re/.

2 Extracting Meaningful Biological Relations

Methods for extracting biological information from the scientific literature have improved considerably [10, 11]. Entity recognition tools use specialised lexical resources, such as dictionaries and ontologies and high-quality training resources, i.e., annotated corpora. Relation extraction approaches range from simple statistical heuristics (e.g. by considering co-occurrences of search terms or estimating term frequency distributions) to combined syntax and

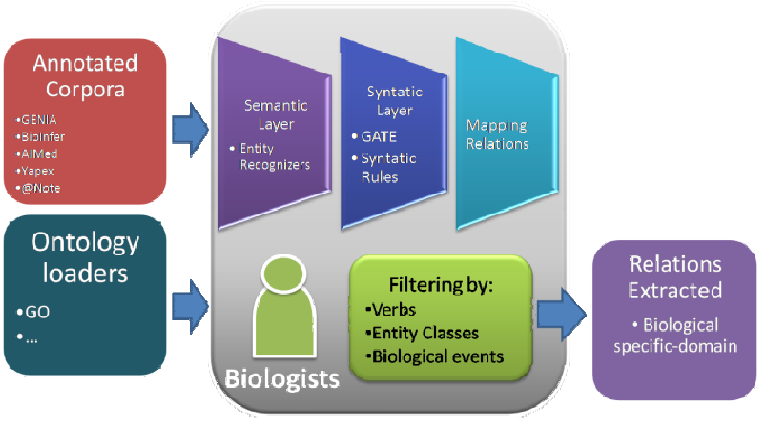

Figure 1. General relation extraction workflow.

semantic sentence parsing using NLP techniques [12-14].

The next subsections will describe our software development efforts, highlighting biologist intervention throughout the workflow (Figure 1).

2.1 Entity Recognition

Our tool assumes that entity recognition has already been performed by available entity recognisers and thus, we implemented loaders for common annotation schemas such as GENIA [3,4], BioInfer [5], AIMed [6], Yapex [7] and @Note [8]. Internally, we consider five major entity classes, namely: genes, proteins, compounds, organisms and others. Loaders follow the original schema directives while performing class mapping, i.e., we do not alter the original annotations in any way. Class grouping simplifies analysis, but original classes are present and can be looked into.

2.2 Syntactic Processing

Our tool is prepared for the inclusion of different NLP tools since different problems can benefit from different general and specialised approaches. So far, we implemented a general relation extraction model using GATE language engineering software [9]. GATE grants shallow linguistic processing, such as tokenisation, sentence splitting, Part-of-Speech (POS) tagging and lemmatisation. Also, it enables noun phrase and verb phrase grouping.

A set of candidate relations indexed by their verbal forms (verb lemmas and actual verb conjugations) is outputted. Each potential relation is characterised in terms of left and right hand entities, polarity (negative/positive statement) and directionality (triggering entity vs. affected entity). For example, in Figure 2 the entities "Y292" and "ZAP-70" (previously classified as "amino acid monomer" and "protein molecule", respectively) are related by the verb form "regulates" and the adverbial form "negatively" points out the polarity of the relation, i.e., states that this is a negative form of regulation.

2.3 Analysis of Relation Properties

After performing relation extraction, the biologist is presented with relation and entity and entity class report views. Each view supports cardinality, directionality and polarity filters that automatically zoom into the particular scope of analysis the biologist is interested in (Table 1). The analysis of cardinality is centred in entity and entity class annotations. Left and right-hand frequencies identify the entities (or entity classes) most commonly associated to given relations. For example, the biologist may look into regulatory relations involving multiple regulatory genes or reactions with a single metabolite.

Relations may also be characterised in terms of direction. Undirected relations (e.g. protein-protein interactions) and directed relations (e.g. regulatory relations) can be studied separately and directed relations can be further inspected. Determining the role of each entity in the relation, namely cause-effect characterisation, may bring interesting results in many biological problems.

Polarity studies may also provide important evidences to the biologist [15]. Besides tracking down words with an inherently negative meaning (e.g. absent, fail, lack and exclude), the tool identifies affixal negation forms (e.g. "inactivate" and "deactivate" are affix negation forms of the verb "activate"), negative determiner forms (e.g. "No interaction was identified" or "Nothing was identified") and direct negation forms (e.g. "not activate" and "inactivate").

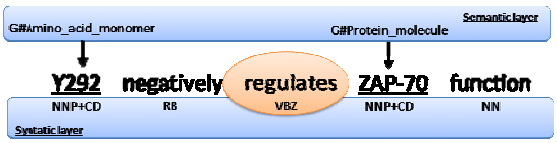

Figure 2. Examples of syntax-semantic combined annotations.

Table 1. Examples of relations and associated properties extracted from the GENIA corpus.

Relation Properties	Example
Cardinality	**One-to-one "**... T lymphocytes activate NF-kappa B..." (PMID: 94354848) **One-to-many/many-to-one "**... Th2 cells **produce** IL-4, OL-5, IL-6, IL-10, and IL-3..." (PMID: 94354848) **Many-to-many "**... anti-CD28 mAb or CHO cells **expressing** the CD28 ligands CD80 and CD86..." (PMID: 99008506)
Directionality	**Undirected "**... the STAT1 alpha protein bound to the Fc gamma RIC GIRE ..." (PMID: 96032864) **Left-to-right "**... macrophages express detectable HIV proteins..." (PMID: 91218850) **Right-to-left "**... Akt/PKB is regulated by Ras signalling pathways..." (PMID: 99010726)
Polarity	**Adverbial negation** "... response in T cells that **was not accompanied** by measurable IL-2 production ..." (PMID: 97025433) **Affixal negation "**... human T lymphocyte cultures **are unable to undergo** proliferation..." (PMID: 94172207); "...muE3 and muB **inactivated** the mu enhancer in S194 plasma cells..." (PMID: 96315681) **Emphatic negation "**... **no evidence could be found** that the virus ever circularizes..." (PMID: 95266275) **Negative nominals "**... Lymphocytes from CML patients **lack** a 47 kDa factor..." (PMID: 97119289)

2.4 Ontology-based Semantic Mapping

Biologists are used to analyse domain-specific concepts rather than verbal-indexed relations. For instance, instead of looking into "regulate", "inhibit", "activate" relations separately, the biologist is probably interested in establishing that all three lemmas are associated to the concept "regulation of biological process".

Ontology-based mapping helps contextualise relevant relations, characterising and classifying them according to the domain under study. Thus, we included into our tool a graphical interface where biologists are presented with an ontological hierarchy and the extracted relations indexed by verb lemma. Based on his domain of expertise and the different verb lemmas extracted from the corpus, the biologist evaluates which relations are interesting for his analysis. A simple drag and drop of verb lemmas into ontological concepts is just what it takes for the biologist to contextualise relations. The biologist is responsible for the number and kind of associated relations and the assessments taken in each association. Maybe two experts in a given domain will not perform the same mapping, reflecting natural discrepancies of judgment and interests. This is exactly what we are looking for with this interface. Biologists have the possibility to make their own evaluation without being bound or constrained to any kind of rules or directives of consensus. Actually, the only current limitation imposed to manual curation is that all verb forms of a given lemma are to be associated to the same ontological concept. Even so, this restriction can be eliminated in future versions if it is considered to be useful in some scenarios.

We chose to include Gene Ontology (GO) as primary ontological resource due to its broad-scope and extensive annotation. It encompasses three ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner [16]. Hence, it may be of assistance in virtually any biological scenario.

2.5 System Evaluation

The system supports benchmarking corpora evaluation, assisting the biologist to assess the performance of different text mining workflows on his/her particular domain. High-quality annotated corpora are required in this process. So far, we provide loaders for the GENIA Events [3] and BioInfer [5] benchmarking corpora.

Additionally, the biologist is able to manually curate the set of candidate relations, eliminating irrelevant relations, refining relation annotation (e.g. extending incomplete verbal forms) and annotating new relations (e.g. by inspecting entities that have not been associated with any relation).

3 Conclusions

Ultimately, literature mining aims at hypothesis generation and biological discovery in any given domain. This aim is quite bold, encouraging tight collaboration between biologists and text miners. Text miners provide for automated text processing techniques whereas biologists filter, contextualise and analyse outputs.

Gathering together tools and expert knowledge is not easy, especially if tools are to be domain-independent. This was the main motivation for our work. Our software addresses relation extraction based on the abilities of common entity recognisers and syntactic parsers. Domain-specific analysis is enabled by expert-driven ontological mapping of extracted relations to meaningful biological relations. Its primary contributions are as follows:

- a pre-processing module capable of loading corpora for common annotation schemas (e.g. GENIA, BioInfer, AIMed, Yapex and @Note);

- a NLP module which supports common (such as tokenisation, sentence splitting and POS tagging) as well as advanced (such as verb and noun phrase chunking) text processing;
- a user-friendly interaction platform that allows the expert to work over general outputs, studying them within the scope of a given domain.

By contextualising outputs, biologists get the best out of general relation analysis, since studies on co-occurrence, directionality and polarity are more comprehensible.
Future work includes loaders for additional annotation schemas and more NLP techniques. Relation processing, namely the identification of special cases of negation and directionality, is to be refined. Working with other available ontologies or even specifying new ones according to expert directives is also being considered.
The software is open-source and is freely available at http://sysbio.di.uminho.pt/anote/re/.

Acknowledgments

The work of Sónia Carneiro is supported by a PhD grant from the Fundação para a Ciência e Tecnologia (ref. SFRH/BD/22863/2005). The work of Rafael Carreira is supported by the SYSINBIO coordination and support action (call FP7-KBBE-2007-1).

References

1. P. M. Roberts, "Mining literature for systems biology," *Briefings in Bioinformatics*, vol. 7, no. 4. pp.399-406, Dec., 2006.
2. K. B. Cohen and L. Hunter, "Getting started in text mining," *Plos Computational Biology*, vol. 4, no. 1, Jan., 2008.
3. J. D. Kim, T. Ohta, and J. Tsujii, "Corpus annotation for mining biomedical events from literature," *BMC Bioinformatics*, vol. 9, Jan., 2008.
4. J. D. Kim, T. Ohta, Y. Tateisi et al., "GENIA corpus-semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19 Suppl 1. pp.i180-i182, 2003.
5. S. Pyysalo, F. Ginter, J. Heimonen et al., "BioInfer: a corpus for information extraction in the biomedical domain," *BMC Bioinformatics*, vol. 8, Feb., 2007.
6. R. Bunescu, R. F. Ge, R. J. Kate et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2. pp.139-155, Feb., 2005.
7. K. Franzen, G. Eriksson, F. Olsson et al., "Protein names and how to find them," *International Journal of Medical Informatics*, vol. 67, no. 1-3. pp.49-61, Dec., 2002.
8. A. Lourenço, R. Carreira, S. Carneiro et al., "@Note: A Workbench for Biomedical Text Mining". Journal of Biomedical Informatics [Accepted]
9. H. Cunningham, D. Maynard, K. Bontcheva et al., "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications." *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)* . 2002.
10. H. J. Dai, J. Y. W. Lin, C. H. Huang et al., "A Survey of State of the Art Biomedical Text Mining Techniques for Semantic Analysis." *IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing (SUTC '08).* pp. 410-417. 2008.
11. A. Skusa, A. Ruegg, and A. Kohler, "Extraction of biological interaction networks from scientific literature," *Briefings in Bioinformatics*, vol. 6, no. 3. pp.263-276, Sept., 2005.
12. F. Rinaldi, G. Schneider, K. Kaljurand et al., "Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach," *Artificial Intelligence in Medicine*, vol. 39, no. 2. pp.127-136, Feb., 2007.
13. J. D. Wren, "Extending the mutual information measure to rank inferred literature relationships," *BMC Bioinformatics*, vol. 5, Oct., 2004.
14. J. Atkinson and A. Rivas, "Discovering Novel Causal Patterns From Biomedical Natural-Language Texts Using Bayesian Nets," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 6. pp.714-722, Nov., 2008.
15. O. Sanchez-Graillet and M. Poesio, "Negation of protein-protein interactions: analysis and extraction," *Bioinformatics*, vol. 23, no. 13. pp.I424-I432, July, 2007.
16. D. P. Hill, B. Smith, M. S. McAndrews-Hill et al., "Gene Ontology annotations: what they mean and where they come from," *BMC Bioinformatics*, vol. 9, 2008.