Text Mining bridging the gap between knowledge and text (Extended Abstract)

Sophia Ananiadou
NaCTeM, University of Manchecter
kis@imet.ac.ru

Useful pathway models require a complete and accurate representation of the system, which requires that all relevant molecular species are captured, together with their physical interactions and chemical reactions. Pathway model reconstruction is currently largely carried out manually by domain experts, who must carefully read the scientific literature, in order to retrieve, evaluate and interpret and distil relevant fine-grained statements. Moreover, due to the proliferation of scientific databases and ontologies, discovery of previously unknown knowledge demands that scientists take into account information from many different resources, covering different levels of contextual information (e.g., degree of confidence or certainty expressed towards a finding). Thus, given the high complexity mechanisms involved in pathway models, whose detailed description can only be derived from analysis of heterogeneous, fragmented and incomplete sources, reconstructing pathway models is a slow, difficult and laborious process. Accordingly, there is a need to develop methods that help experts to make sense of the continuously growing body of literature, in order to increase the speed and reliability of knowledge discovery.

In response to the above, text mining (TM) aims to automate the above process, by finding relations (such as interactions) that hold between concepts of different types (e.g., genes/proteins, chemical compounds, metabolites, subcellular components, anatomical entities, organisms, cell lines, strains, diseases). A large number of TM methods aim to extract simple binary relations from e.g., A binds B. This is mainly achieved by focusing on textual co-occurrences, using bag-of-words approaches, analysis of controlled vocabulary metadata, and other shallow techniques. However, these approaches have several disadvantages, including the identification of many false positive relations. Additionally, they fail to take into account contextual information about relations, e.g., the cellular context of a signaling event, such as cell type and localization.

In contrast, our work involves the development of more sophisticated TM techniques to extract events, which encapsulate typed n-ary relationships, i.e., interactions between any number of concepts. Events are able to capture detailed information about mechanisms of biological pertinence, e.g.,, reactions such as negative regulation, phosphorylation, carboxylation), by linking together interacting participants, which play specific roles (e.g., modifier, reactant, product, cause, location). As such, they are able to encode several types of contextual information, that are frequently missing when only binary relations are considered.

Consider an intuitive example from the literature to explain our goal: The results suggest that the narL gene product activates the nitrate reductase operon. (PMID: 3035558). This sentence provides interpretative information about the reaction between the narL gene product and the nitrate reductase operon, namely that the information stated in based on an analysis/interpretation of experimental results, and that there is a certain amount of speculation expressed towards the reaction (according to the use of the verb suggest, rather than a more definite verb, such as demonstrate). Next, consider a more complex example: The analysis showed that IEXC29S was unable to significantly transactivate the c-sis/PDGF-B promoter. Whilst a conventional TM analysis to find binary relationships would simply discover that some type of interaction occurs between IEXC29S and c-sis/PDFG-B, a more detailed contextual analysis would allow the construction of a representation that encodes the complex details of the interaction, e.g., that the information is stated based on an experimental analysis, and that the interaction has been shown to occur with a low level of intensity. .

In order to extract such complex events automatically, we have developed a pipeline-based event extraction system, EventMine [1], which employs a series of classifier modules to capture core event elements: detection of triggers (words or phrases that characterise the event; typically verbs or their nominalisations ,detection of edges (finding links between pairs of concepts), and complex event detection (combining multiple edges of complex n-ary relations).

EventMine utilises a rich set of features including those obtained from dependency parse trees supplied by the GENIA Dependency Parser [2], as well as from predicate-argument structures determined by Enju [3], which has been adapted for application to biomedical text. EventMine is capable of extracting interactions across different sentences, owing to its capability to incorporate results from a pre-executed coreference resolution method [4]. In this way, event participants

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

which are semantically empty (e.g., expressions such as it, that) are resolved to their referents and thus become more informative. In addition, the system can be adapted to different tasks without the need for task-specific tuning [5]. Finally, EventMine also facilitates the extraction of interpretative context by detecting various event attributes, e.g., polarity, certainty, manner, knowledge type and source [6]. As with its other classifier modules, EventMine uses SVMs for this task, facilitated through its training on the GENIA Meta-knowledge corpus [7].

The automatic extraction of events from biomedical text has a broad range of applications, which include not only support for the creation and annotation of pathways [8], but also automatic population/enrichment of databases and semantic search systems. To develop systems that are customized for different tasks such as the above, a text mining infrastructure is needed, which is able to support the curation and maintenance of pathways, sharing and re-using of knowledge distributed over thousands of scientific publications and monitoring of recent publications is needed to maintain relevance. To foster adaptability of TM solutions, we are using our UIMA based Argo platform [9], which enables the development of highly customisable solutions in the form of reconfigurable modular text mining pipelines (workflows). Apart from supporting the straightforward integration of application-specific components, reconfigurable workflows allow for discovery of optimal solutions [10] owing to their interchangeable underlying components. For components to be interoperable (i.e., for one component to build on the text annotations created by another), the outputs of a predecessor component must be type-compatible with the inputs expected by a successor. In Argo, this is ensured by mechanisms that support mapping between similar semantic types and conversion of annotations [11].

Argo has supported the development of systems such as PathText2[1], an integrated search system that links biological pathways with supporting knowledge in the literature [8]. It reads formal pathway models (represented in the Systems Biology Markup Language (SBML) and converts them into queries that are submitted to three semantic search systems operating over MEDLINE, i.e., KLEIO, which improves and expands on standard literature querying with semantic categories and facetted search, FACTA+ (see below) and MEDIE (http://www.nactem.ac.uk/medie/), which extracts events. MEDIE has been found to achieve the highest hit ratio, which demonstrates the superiority of events for finding relevant interactions.

FACTA+[2] [12] discovers hidden, previously unknown associations between both concepts and complex events from the literature (such as Gene expression, Positive regulation, Binding, Regulation, etc.). Such associations can often only be found by linking information that may be dispersed across many documents, and thus which might be missed using manual search methods. This facilitates hypothesis generation, which is directly relevant to pathway construction. FACTA+ approaches the problem of automatic discovery of useful hypotheses by combining two (or more) known associations, i.e., if concept X is associated with concept Y and concept Y is associated with concept Z, then the potential indirect association between X and Z is considered as a useful hypothesis unless there is already a known association between X and Z. FACTA+ supports the discovery of indirect associations based not only on concepts but also on complex events such as Gene expression, Positive regulation, Binding, Regulation, etc.

Advanced TM methods such as those described here support pathway curation, validation and maintenance. Their employment yields improved coverage, faster acquisition and throughput, combined with easier identification and normalisation of duplicates, greater consistency, completeness and accuracy in description, and reduced curator burden. This helps to free experts from mundane and tedious tasks while aiding with more intellectually challenging ones.

References

[1] Miwa M, Saetre R, Kim JD, Tsujii J. Event extraction with complex event classification using rich features. J Bioinform Comput Biol. 2010;8(1):131-46. Epub 2010/02/26. doi: S0219720010004586 [pii]. PubMed PMID: 20183879.

[2] Sagae K, Tsujii Ji. Dependency parsing and domain adaptation with LR models and parser ensembles. Proceedings of CoNLL 2007 Shared Task; 2007. p. 1044-50.

[3] Miyao Y, Sagae K, Saetre R, Matsuzaki T, Tsujii J. Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics. 2009;25(3):394-400. PubMed PMID: 19073593.

[4] Miwa M, Thompson P, Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. Bioinformatics. 2012;28(13):1759-65. doi: 10.1093/bioinformatics/bts237.

[5] Miwa M, Ananiadou S. Adaptable, high recall, event extraction system with minimal configuration. BMC Bioinformatics. 2015;16(10):1-11. doi: 10.1186/1471-2105-16-s10-s7.

[6] Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S. Extracting semantically enriched events from biomedical literature. BMC Bioinformatics. 2012;13(1):108.

[7] Thompson P, Nawaz R, McNaught J, Ananiadou S. Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinformatics. 2011;12:393.

[1] http://www.nactem.ac.uk/pathtext2/demo/

[2] http://www.nactem.ac.uk/facta/