Towards an Automated Analysis of Biomedical Abstracts

Barbara Gawronska, Björn Erlendsson, and Björn Olsson

School of Humanities and Informatics, University of Skövde
Box 408
541 28 Skövde, Sweden
barbara.gawronska@his.se,
bjorn.erlendsson@his.se,
bjorn.olsson@his.se

Abstract. An essential part of bioinformatic research concerns the iterative process of validating hypotheses by analyzing facts stored in databases and in published literature. This process can be enhanced by language technology methods, in particular by automatic text understanding. Since it is becoming increasingly difficult to keep up with the vast number of scientific articles being published, there is a need for more easily accessible representations of the current knowledge. The goal of the research described in this paper is to develop a system aimed to support the large-scale research on metabolic and regulatory pathways by extracting relations between biological objects from descriptions found in literature. We present and evaluate the procedures for semantico-syntactic tagging, dividing the text into parts concerning previous research and current research, syntactic parsing, and transformation of syntactic trees into logical representations similar to the pathway graphs utilized in the Kyoto Encyclopaedia of Genes and Genomes.

1 Background and Aim

Text mining has many applications in the area of bioinformatics, where computerized tools are used to analyze data concerning molecular biological objects (genes, proteins, gene regulation pathways, cells, etc) in order to derive new biological insights [1], [2]. The aim of the research described in this paper is to develop a system that applies automated text analysis to support the large-scale analysis of metabolic and regulatory pathways by deriving relevant relations from textual descriptions found in the literature. The need for such a system arises from the fact that molecular biologists today need efficient computer-based tools to navigate the huge amount of knowledge that has been generated over the years and documented in published papers. Since it is becoming increasingly difficult to keep up with the vast number of scientific articles being published, there is a need for more easily accessible representations of the current knowledge. The KEGG pathway database [3] is one example of such an effort to systematically collect the current knowledge on molecular interaction networks in biological processes. Building knowledge bases manually, however, is extremely time-consuming, since each pathway map in KEGG is based on findings from a large number of experiments which have been reported in separate research articles.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 50 – 65, 2006.

Although databases such as KEGG provide easily accessible sources of knowledge for the user, they require enormous amounts of work to build, maintain and keep up-to-date. Therefore, the long-term aim of the research presented here is to provide a semi-automated method of deriving pathway maps using a text corpus as input. As indicated in the overview in Figure 1, we view text analysis as one component in a system that derives pathways from biomedical texts selected from PubMed, using lexical databases and a grammar-based in-depth analysis.

Automated text analysis offers support for the process of structuring knowledge, provided it is conducted using in-depth text comprehension methods. Many of the text mining efforts in bioinformatics, however, have been based on using only statistics regarding co-occurrence of terms [4-11]. As pointed out in [7], this frequent use of simple co-occurrence owes its popularity to the fact that it is easy to implement and allows efficient processing of huge amounts of texts. Such text retrieval and text mining devices can inform the researcher that there seems to be some relation between e.g. a gene and a protein, but in most cases they do not specify what kind of relation it is.

Another line of research is to use pattern- and template-based approaches [12-16]. For example, [17] used a protein name dictionary together with surface clues on word patterns and simple part-of-speech rules to predict protein interactions. In a similar effort, [18] developed a method (BioNLP) based on pattern-matching, which searches for sentences matching a set of rules describing selected functions carried out by proteins. The work in [19] represents a hybrid approach (a stochastic word tagger is combined with rule-based semantic and syntactic analysis). In general, there has been a shift of focus in recent years towards methods which make use of rules and grammars. Examples can be found both in bioinformatics [6], [20] and biomedical information extraction [21], as well as in other domains [22]. As pointed out in [13], a restriction common to most relation extraction models is the lack of ability to extract more than one relation per sentence. Another shortcoming is that relations not expressed by verbs but by, e.g., nouns or participles, are normally omitted.

Among the on-line available information extraction tools, MedScan [6], [20] includes an ambitious attempt to extract positive and negative regulation relations from texts. The developers of the system stress the importance of analyzing subordinated clauses and taking modality into account. We tested the recently released version of MedScan (available at http://www.ariadnegenomics.com/products/medscan.html) on a corpus of 40 biomedical abstracts and found that - although the precision has improved compared to the earlier version - the system is still not reliable enough. As pointed out in [6] the coverage is low. In our corpus of 40 abstracts, only 19 biological relations were found. Out of these 19 relations, it was found upon manual inspection that at most 9 had been correctly extracted. Errors were due mainly to insufficient grammatical analysis. Especially subordinated clauses, ellipsis, appositional constructions, and coordination caused problems. Also long noun sequences caused evident difficulties, and the distribution of the extractions was uneven. Biological relations were extracted from 8 abstracts, while in the remaining 32 no relations were found, although manual inspection revealed that most of these abstracts mentioned several relevant relations. Furthermore, it seems that the system identifies biological objects only if their names are present in the specialized lexicon/ontology it has access to.