Improving Text Mining with Controlled Natural Language: A Case Study for Protein Interactions

Tobias Kuhn[1,2], Loïc Royer[1], Norbert E. Fuchs[2], and Michael Schröder[1]

[1]Biotechnological Center, TU Dresden, Germany
{loic.royer, michael.schroeder}@biotec.tu-dresden.de,
http://www.biotec.tu-dresden.de/schroeder
[2]Department of Informatics, University of Zurich, Switzerland
{tkuhn, fuchs}@ifi.unizh.ch,
http://www.ifi.unizh.ch/attempto

Abstract. Linking the biomedical literature to other data resources is notoriously difficult and requires text mining. Text mining aims to automatically extract facts from literature. Since authors write in natural language, text mining is a great natural language processing challenge, which is far from being solved. We propose an alternative: If authors and editors summarize the main facts in a controlled natural language, text mining will become easier and more powerful. To demonstrate this approach, we use the language Attempto Controlled English (ACE). We define a simple model to capture the main aspects of protein interactions. To evaluate our approach, we collected a dataset of 459 paragraph headings about protein interaction from literature. 56% of these headings can be represented exactly in ACE and another 23% partially. These results indicate that our approach is feasible.

1 Introduction

In this paper we introduce a new paradigm of how to make knowledge of scientific papers accessible by computers. We focus on the fields of life sciences – particular biology – but our approach could be used in other fields as well.

Our approach consists of letting authors express their scientific results in a formal summary that could be an integral part of the papers they publish. We argue that it is more reasonable to let the authors formalize their own results, instead of trying to extract these results from the articles.

This section explains our motivation, introduces the language Attempto Controlled English (ACE) and compares it with other knowledge representation languages. Section 2 shows how ACE is used to build an ontology for protein interactions. In Sect. 3 we use this ontology as foundation for the expression of scientific results and we show how 89 selected articles could have been summarized in ACE. Section 4 shows the benefits of our approach and Sect. 5, finally, gives a short outlook.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 66–81, 2006.

1.1 Motivation

Biomedical scientists are challenged by an ever-increasing amount of scientific papers. The indexing service *PubMed*[1] shows the huge quantity of literature that the scientists have to face. It contains at the moment 16 million articles and grows every year by over 600'000 articles. All these biomedical articles are written in natural language. That means that we cannot easily process them with computers. But, facing the quantity of literature, it is clear that we need computational support in order to manage the contained knowledge.

In the last years, *text mining* and *information extraction* – which build both upon natural language processing (NLP) – gained an increasing interest in biomedical sciences. They aim to extract some kind of formal knowledge from natural language texts, which is generally considered a very demanding task. Even the basic problem of *named entity recognition*, that aims to identify named entities (e.g. protein names) in natural texts, is far from being solved. Other major aspects of text mining are the extraction of relationships (e.g. protein interactions), the automatic classification of texts, and the generation of new hypotheses on the basis of the available literature [3]. The *BioCreAtIvE* contest [21] nicely shows, that even sophisticated tools for text mining have a considerable lack of precision and recall: For a simple "named entity recognition"-task the precision ranged up to 86% and the recall was at most 84%. Another attempt is described in [4]: Information about protein-interactions was extracted from a data set of 1.2 million sentences that were taken from biomedical abstracts. They achieved a precision of 91%, but with a poor recall of only 21%. We recommend [3] and [12] for a more comprehensive overview of the "accomplishments and challenges" of text mining.

As a first step towards a better management of biomedical literature, controlled vocabularies like *MeSH*[2] and the *Gene Ontology*[3] have been created. They serve to classify biomedical publications and to link them to other resources. *GoPubMed*[4], for example, is a search engine that connects the abstracts from PubMed with the formal structure of the Gene Ontology. Thus a researcher can exploit the Gene Ontology for the search of relevant literature. Such tools are very valuable for scientists and there has been a notable progress in the last years, but it will never be possible to extract all the information correctly. There is inherent ambiguity and vagueness in natural language that prevents its perfect processing by computers.

For this reason we present an alternative approach: The authors of scientific articles formally summarize their own results. Such formal summaries are added to the articles which makes them processable by computers. This requires a formal language that on the one hand is easy to learn and understand, and on the other hand is expressive enough to represent even complicated scientific results. It is clear that this approach is not applicable for papers that have been

[1] `http://www.pubmed.gov`
[2] `http://www.nlm.nih.gov/mesh/meshhome.html`
[3] `http://www.geneontology.org`
[4] See [5] and `http://www.gopubmed.org`