FEBS Letters 582 (2008) 1178–1181

Minireview

A text-mining perspective on the requirements for electronically annotated abstracts

Florian Leitner, Alfonso Valencia*

Structural Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

ailable online 6 March 2008

Abstract **We propose that the combination of human expertise and automatic text-mining systems can be used to create a first generation of electronically annotated information (EAI) that can be added to journal abstracts and that is directly related to the information in the corresponding text. The first experiments have concentrated on the annotation of gene/protein names and those of organisms, as these are the best resolved problems. A second generation of systems could then attempt to address the problems of annotating protein interactions and protein/gene functions, a more difficult task for text-mining systems. EAI will permit easier categorization of this information, it will help in the evaluation of papers for their curation in databases, and it will be invaluable for maintaining the links between the information in databases and the facts described in text. Additionally, it will contribute to the efforts towards completing database information and creating collections of annotated text that can be used to train new generations of text-mining systems. The recent introduction of the first meta-server for the annotation of biological text, with the possibility of collecting annotations from available text-mining systems, adds credibility to the technical feasibility of this proposal.**

Keywords: Information extraction; Article annotation; Text mining; Journal annotation pipeline; Review; Perspectives; Electronically annotated information

1. Introduction

The continuous growth in the number of research articles and the corresponding data stored in online repositories require better connections to be established between scientific articles, annotations and data [1]. Coherent annotation will enhance the possibility of locating and comparing articles, while database links will permit the unambiguous recovery of information [2]. Depending on an article's content and type, such annotations may include the names of genes or proteins discussed in the article and the database identifiers where the sequence data are stored. These annotations may also contain indicators of concept ontology for diseases, molecular functions, cellular locations or experimental methods [3]. They may even cover structured annotations of figures and tables [4].

Although it would be desirable for these annotations to be exclusively made automatically through natural language understanding (NLU), the intrinsic difficulty of natural language makes automation very difficult [5] and currently, only semi-automated approaches can realistically be considered in practice [6]. This review presents the state of the art in information extraction (IE) systems that could be used to annotate biological text and to build annotated abstracts (for current reviews on IE tools, see [7–9]). The potential requirements for such annotation process will be examined, highlighting possible approaches for the annotation of articles and the creation of summaries (electronically annotated summaries).

2. Current state of the art and opinions

Automated extraction of classified content (IE) is an important area of biological text mining [10]. The status of current IE systems is currently being evaluated in the light of the BioCreative challenge. BioCreative (critical assessment for information extraction in biology) systematically assesses the systems currently available with the help of experts in specialized databases [11,12]. The evaluation includes methods for the ranking of documents according to their biological relevance (document classification task), the identification of biological entities (i.e., protein and gene names, known as named entity recognition, NER), as well as the more complex detection of relationships between entities (i.e., protein interactions, protein and biological functions).

The entity detection task can be divided into the identification of gene and protein name mentions in the text on the one hand, and into the assignment of unique database identifiers to the gene and protein names on the other (a process known as normalization). The BioCreative results show that the best available NER systems are able to correctly detect almost 90% of the names (with 88% precision and 86% recall for gene mention detection, and 83% precision and 79% recall for the normalization task) [13,14]. However, it can be assumed that the identification of other types of entities, such as cell types, chemicals, diseases, and others will present additional difficulties.

The results on the more complex second task of associating genes/proteins to gene ontology terms analyzed in BioCreative I [15], or of detecting protein interactions analyzed in BioCreative II [16], were far less positive. These results clearly demonstrated the need of intensive human intervention to complement the automated results, especially when thinking about

*Corresponding author.
E-mail address: valencia@cnio.es (A. Valencia).

Abbreviations: BCMS, BioCreative MetaServer; EAI, electronically annotated information; IE, information extraction; NER, named entity recognition; NLP, natural language processing; NLU, natural language understanding

0014-5793/$34.00 © 2008 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2008.02.072

real-world applications. Generally, relationship extraction requires linking entities with their associated concepts, for example, linking a gene name to a disease while annotating the mutations associated, or coupling two interacting proteins with a specific type of interaction [17,18]. Part of the difficulty in this process is related to the initial complexity of the recognition and normalization of names. This problem includes both the identification of the protein names and the corresponding species, and the additional problem of the semantic identification of the relationships [19]. It is worth mentioning that even human experts often disagree about the detection and classification of entity relationships, as revealed through a variety of inter-annotator agreement exercises [20]. Indeed, it is not uncommon for database annotators to request additional information from the authors after having studied a publication.

BioCreative has also shown that applying natural language processing (NLP) techniques to full-text articles is in general much more difficult than processing abstracts alone (as commonly done in text-mining publications). For example, a problem in full-text articles may lie in the significance of the information in different parts of the text, which becomes a relevant and additional burden.

In summary, while the first two IE objectives (document classification and NER) are problems for which current IE systems already provide sufficiently robust online solutions, the extraction of relationships is still a task for which IE tools require additional development [21].

BioCreative and other related experiments have also consistently identified the demand for creating sufficiently large collections of annotated articles (termed corpora) that can be used to train and test text-mining systems [22]. Indeed, all the IE and most of the NLP methods are based on learning features from previously annotated text, even though an interesting number of methods based on training with semi-annotated text have been proposed. Pioneering efforts to create such corpora were started by the GENIA initiative [23], but they also include the BioCreative collections as well as others [24]. Nevertheless, insufficiently large collections are currently a key limitation that undermines the development of sophisticated text-mining strategies. This makes biology different from other fields where these collections are readily available.

A major issue in biological text mining and clearly a keystone for future developments is the need to generate and use consistent annotations. These must follow common standards in well-structured representations, and very importantly they must be linked to the corresponding text sources. Interestingly, and for completely other reasons discussed elsewhere [25], the traditional separation between database records and journal entries is vanishing. The sum of these developments is the generation of an environment in which the publication of papers is becoming more directly related with the deposition of the basic information in databases. Hence, the need for the annotation of the corresponding text with basic links to the databases, and to combine automatic annotation with human expert (and if possible, author made) annotations. Possible scenarios that have been discussed include (see [26–28]):

- *Allowing authors to freely decide on the annotations*. In this case, it will difficult to maintain consistency that would largely restrict the possible application of automatic IE and database-related tools.
- *Letting authors choose concepts from collections of controlled vocabularies (e.g., gene ontology (GO) terms)*. This generates the obvious difficulty of obtaining consistent annotations from the authors not necessarily aware of how to use the ontology. Moreover, it should be born in mind that training a GO annotator requires several months of work in close collaboration with experts.
- *Semi-automated systems that would pre-filter possibly relevant terms from collections of controlled vocabulary and offer the results to human experts for validation.* This type of system can bridge the need between annotation, consistency and annotator expertise.

An important additional issue is the technical feasibility of annotating text and the accessibility of the systems/software to achieve this. To date, there is no single stable/open system for the full annotation of papers, and most of the current publications refer to complex systems combining sets of text-mining tools developed internally by various research groups. To address this situation, BioCreative II has developed a meta-server able to collect and unify results from distributed systems working in the classification and annotation of MEDLINE abstracts [29]. The BioCreative MetaServer (BCMS) is an online platform using web services to collate annotations, including gene/protein names and normalizations, and document classifications of protein interactions and taxa. In the current implementation, results from 12 different systems are included in the meta-server. While a complete article annotation system would have to go beyond this functionality, BCMS can be viewed as a first demonstration of the viability of such a pipeline and/or as an initial prototype.

3. Towards electronically annotated articles

To design a system for semi-automated article annotation the potential scenarios in which they will be used must be taken into account. The most prominent situations, that are indeed associated one with the other, are

1. The journals that wish to provide their users with access to their articles through a structured and hierarchical interface, sub-categorizing the publications based on the annotations.
2. Database annotators that wish to easily retrieve articles relevant to their data repository, the annotations associated greatly reducing time consuming tasks such as normalizing gene and protein names, mapping controlled vocabulary, and maintaining existing records.
3. Biological text mining could concentrate on the more challenging tasks of uncovering complex and implicit information, using the curated annotations as their starting point.
4. Most importantly, providing researchers with direct unambiguous access to raw data sources and the capacity to retrieve relevant articles with high precision and recall. For example, to find specific methods and techniques, or to monitor current progress in their field of interest. The annotations, which can be regarded as summaries of an article, will permit the rapid assessment of publication quality and relevance.

4. Proposal for the requirements of an interactive electronic annotation system

Taking into consideration the possible scenarios of usage and the basic categories of annotation types described above, our proposal for such a system would be (see Fig. 1)