BioExcom: Automatic Annotation and categorization of speculative sentences in biological literature by a Contextual Exploration processing

Julien Desclés
LaLIC Université Paris-Sorbonne
Maison de la Recherche
28 rue Serpente, 75006 Paris
descles2@yahoo.fr

Motasem Alrahabi
LaLIC Université Paris-Sorbonne
Maison de la Recherche
28 rue Serpente, 75006 Paris
motasem.alrahabi@gmail.com

Jean-Pierre Desclés
LaLIC Université Paris-Sorbonne
Maison de la Recherche
28 rue Serpente, 75006 Paris
jean-pierre.descles@paris.sorbonne.fr

Abstract

Biological research papers are replete with speculative sentences. This paper presents the BioExcom software, an adaptation of EXCOM platform to biology field, which annotates automatically all speculative sentences in full texts papers by the means of the Contextual Exploration processing. This annotation process is based on a concise semantic analysis of the multiple ways of expressing speculation in biology. Furthermore, BioExcom enables to distinguish automatically between prior and new speculations in a biological paper.

Keywords: speculation, biology, contextual exploration, categorization, text mining

1. Introduction

Biological research papers are replete with speculative sentences, also known as *hedges* (Hyland, 1995). For a researcher, it is important to recognize all speculative sentences in a paper or about a given topic. Automatic extraction tools for speculative statements from texts is an emerging field which attempts to meet this need (Light et al., 2004; Wilbur et al., 2006; Di Marco et al., 2006; Kilicoglu and Bergler, 2008; Medlock, 2008; Szarvas et al., 2008). Indeed, biological literature is currently characterized by an extended on-line access and an exponential growth (Rebholz-Schuhmann et al., 2005), which are mostly linked with the development of high-throughput methods and computer science technologies. This huge amount of papers constitutes an extraordinary source of biological facts, knowledge and ideas. However, it is very difficult for a single researcher to keep abreast of all developments (Teufel, 1998; Cohen and Hunter, 2008). To face this challenge, many systems, more elaborate than simple keyword search, have been built (for reviews, see (Hunter and Cohen, 2006; Rzhetsky, 2008)). These systems can combine entity recognition, a statistical processing step, and a rule-based post-processing. However, some authors have reported a lack of performance of these methods for their ability to transform all factual knowledge into a structured formal representation (Rebholz-Schuhmann et al., 2005).

The Contextual Exploration (CE) is a Natural Language Processing method (Desclés, 2006), which constitutes an alternative to classical statistical/machine-learning based technology and to the search for hard-coded linguistic patterns. Furthermore, it does not rely on any preliminary morpho-syntactic analysis. This linguistic method is implemented in a platform, called EXCOM[1], integrating different linguistic resources for text mining (Djioua et al., 2006; Alrahabi and Desclés, 2008). It has been successfully applied to automatic summarization (Blais, 2007), relationships between concepts (Le Priol, 1999), categorization of bibliographic citations (Bertin, 2008) and reported speech (Alrahabi, 2008).

This paper presents BioExcom, an adaptation of EXCOM platform to biological field, which automatically annotates all speculative sentences in biological full text papers by means of the CE processing. The annotation process is based on a concise linguistic analysis of the multiple ways of expressing speculation in biology. Furthermore, BioExcom enables to distinguish automatically between prior and new speculations in a biological paper. We argue that these annotations are useful for biologists, regardless of their domains of interest, to evaluate quickly the content and new output of a paper. We discuss also some possible future applications of speculative sentences extraction and of CE processing in text mining and especially in biology.

2. Task

2.1. Goal

Our goal is to automatically annotate in biological scientific papers all biological speculative sentences (i.e. sentences containing at least one speculative fragment dealing with a biological issue). We consider only sentences with some clear instances of speculative language (the sentence must contain at least one linguistic element expressing speculation). We also want to categorize them into "prior speculation" (speculative sentences cited in the paper, but presented as having been proposed previously) and "new speculation" (speculative sentences presented for the first time in the paper or not explicitly presented as prior speculation). All the examples presented below are sentences from biological literature, found in approximately seventy papers.

[1] For EXploration COntextuelle Multilingue. Find it out at http://www.excom.fr/

2.2. Definition of biological speculation in articles

According to our analysis, it is possible to contrast schematically two types of statements in a biological paper if we consider their degree of certainty:

- *Demonstrated statements*: established facts which are accepted by the scientific community or by the authors of the paper. These can be, for example, biological results, data, observations;
- *Speculations* (non-demonstrated statements): proposals about a biological issue and explicitly presented as not certain in the paper. These can be, for example, hypothesis, interpretations or possible explanations of a fact.

Others types of statements such as deductions, conclusions, argumentation or discussions…, are NOT considered as speculative but as intermediary statements, because they either present things more or less as certain, or they do not make a proposal (the Annotation guidelines is accessible on demand).

3. Automatic annotation of speculative sentences by Contextual Exploration processing

3.1. The Contextual Exploration processing

Contextual Exploration (CE) processing is based on the contextual analysis of linguistic surface markers (no morpho-syntactic parsing) in order to locate discursive expressions used by an author related to a given viewpoint (hypothesis, conclusions, comments, definitions, causal relations, quotations, opinions related to bibliographic references, etc) (Desclés, 2006). This analysis is performed by a linguist with eventually a specialist of a domain, by collecting and categorizing these specific linguistic expressions of a viewpoint.

The linguistic markers of a viewpoint in CE method used for annotating textual segments (which can be a title, a paragraph, a sentence or a clause), are hierarchical: *indicators* and *clues* (expressed into regular expressions). Indicators correspond to linguistic markers (words, discontinuous expressions…), which carry specific information about the studied domain. These linguistic markers can be relatively independent from the authors' style of writing (for instance, "*we present"*, "*in conclusion"*, "*our hypothesis is", "is responsible for"*). However, sometimes the simple presence of an indicator does not permit an annotation of the textual segment in which it appears, because the discourse value of the indicator can change according to the context. Consequently, a more precise annotation of the text may be required. To this end, contextual exploration rules must be applied in order to locate, in the textual context of the indicator, one or more linguistic clues, allowing either the removal of the semantic indecision or a more precise segment annotation. Hence, according to what is specified in the CE rule, the looking for clues can be performed at the right or/and the left of the indicator or even inside the indicator (see an example below).

It is worth noting that the CE processing is different from a classical rule-based system -which consists in searching for specific patterns in a text- in that it makes use of positive and negative clues, the hierarchy between indicators and clues and exploiting the text structure.

3.2. Architecture of the CE engine and overview of text treatment

The overview of our method for automatic extraction of speculative sentences and search for specific speculations is shown in Figure 1. The architecture is based on the EXCOM platform[2] and will not be detailed here, since it has already been described (Alrahabi and Desclés, 2008).

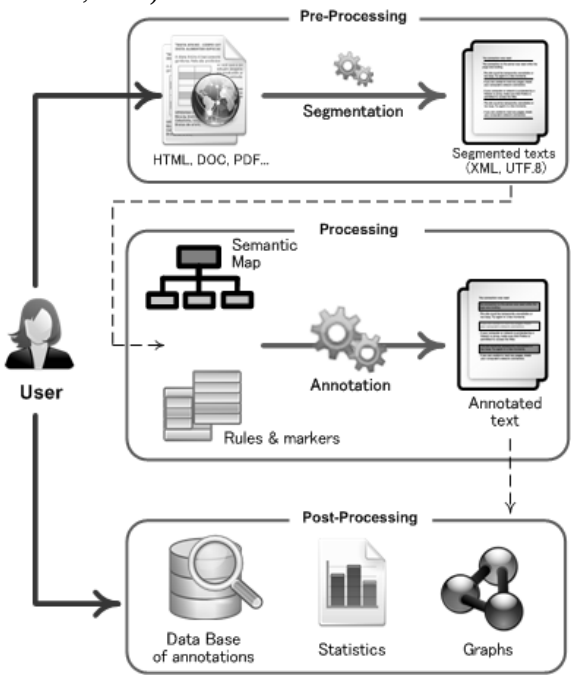

Figure 1: Overview of BioExcom process.

In order to be annotated, all texts must go through the following steps:

1) Automatic segmentation of sentences

In order to split the text into sentences, we use a set of rules, which are based on disambiguation of typographical signs (for example: period, semicolon, question mark, etc.). The input files for the segmentation module are raw text files in UTF-8 encoding, in a given language, and the output files are in the XML DocBook format for articles.

2) Automatic annotation

The core of the platform architecture (Fig. 1) consists of a CE engine that manipulates the indicators and clues as linguistic markers and CE rules associated for annotating linguistic segments. The annotation processing consists in the search for indicators of a given viewpoint in the segment considered. The identification of one indicator calls the associated CE rules. When the conditions of these rules are satisfied (that is, a systematic search for

[2] Principally implemented with Java, JDOM, XML, XLINK, JNLP, etc.

contextual clues, also expressed by regular expressions, in the segment), the CE engine attributes the corresponding annotation to the segment.
In addition, the CE engine is able to establish a hierarchy between rules so as to take into account the fact that some indicators or some rules are more indicative than others. Thus, a sentence will be first analyzed according to a first group of rules, then a second one, and so on. This can also prevent, partly, multiple (possibly contradictory) annotations of sentences.

3) Storage of annotations in a database

Once the texts are annotated, all speculative sentences of the corpus are stored in a "Base of annotations". The annotation scheme of segments contains the following information: the semantic category of the annotation (for example prior or new speculation). The user can choose to consult only prior or new speculations, and can navigate between speculative sentences and their original context by clicking on the sentence: by doing so, he returns to the original annotated paper.

3.3. The linguistic markers of speculation in biological sentences

The linguistic analysis of speculative sentences by the CE processing consists in studying the linguistic markers of speculation at the sentence level. A careful study, carried out by a biologist, on about seventy biological texts[3], has shown that authors use different kinds of specific linguistic markers (or combinations of them) in biological papers, such as:

1) verbs (*to suppose, to suggest, to hypothesize, to propose, to assume*…) ;
2) nouns (*suggestion, hypothesis, speculation*…);
3) adjectives (*convincing, probable, possible, conceivable*…);
4) adverbs (*possibly, probably, perhaps*…) ;
5) modality verbs (*may, might, could*…) ;
6) conjunction (*if, whether, or*…)

3.4. Categorization of speculative sentences

In order to categorize speculative sentences, we look for some specific verbal aspects (for example the passive present perfect applied to specific verbs for "prior speculation"), or the presence of specific constructions (for example, "*we hypothesized*", "*in our theory*" for new speculation) as indicators. If this kind of markers is not available, we look for the presence or the absence of specific clues.
Here are some of the main others clues used for categorizing speculative sentences:

1) *prior speculation*:

- The presence of bibliographic citations in the sentence, as positive clues:
  "*In diatoms, grazing-induced silicification may increase the mechanical resistance of the frustule to copepod mandibules (Hamm et al. 2003).*"
- The presence of others specific words, as positive clues (for example "*recent*" in sentence 1):
  "These recent results with Si and monocots bring not only further support to the theory that Si plays an active role in protecting plants against pathogens, but indicate that this role is not specific to dicots but rather generalized to the plant kingdom."

2) *new speculation*:

- The absence of bibliographic citations in the sentence, as negative clues.
- The presence of other specific words, as positive clues (for example "in this study"):
  "*It is assumed in this study that silicon layers in epidermal cell walls can confer enhanced host resistance to blast*."

3.5 BioExcom implementation

In BioExcom, thirty rules, based on twenty indicator classes (same semantic or grammatical categories), have been implemented and ranked according to seven priorities. We give here one example of implementation of CE rules in the BioExcom system. This is the case of the indicator "*could*", which can be either the past form or the conditional form of "*can*". To take this ambiguity off, we check, in the context of the indicator, the presence or the absence of specific clues expressing conditionality or possibility, such as "*alternatively*" (see the following sentence). Obviously, this method does not allow disambiguating all uses of "*could*", but it correctly recognizes some of them.

"*Alternatively, a soluble Δ9-acyl-ACP desaturase and a membrane-bound Δ9-acyl-lipid desaturase, responsible for the synthesis of 18:1Δ9 and 16:1Δ9, respectively, could co-exist in the plastid of diatoms, similar to the situation found in higher plants.*"

Here is the corresponding simple EC rule, written in a declarative form, used for annotating the sentence as a "*new speculation"*:

"could new speculation" CE rule:
Given P a linguistic segment:
If there is in the before-indicator-context a negative clue from the class "bibliographic references"
And If there is in the after-indicator-context a negative clue from the class "bibliographic references"
If there is in the before-indicator context a positive clue from the class "conditionality"
Or If there is in the after-indicator context a positive clue from class "conditionality"
Then : Give the semantic annotation "New Speculation" to P

4. Evaluation

4.1. Evaluation methodology

As the concept of speculation in biology is not very clear, we need to work with experts giving their judgments about results from BioExcom, which aims to be useful for biologists. We devised an evaluation methodology based on the automatic annotation of new

[3] From very different biological journals (Nature, Science, Plos biology, PNAS, Plant Physiology, Cell…)

and unknown biological articles from different journals by BioExcom and on the random selection of two of them. Between three and five biologists read the version of these papers previously automatically annotated by BioExcom. Before starting the evaluation process, these experts had to read our Annotation guidelines. Then, for each sentence of the annotated articles, they had to say if they were in agreement with the annotation (or the absence of annotation) performed by BioExcom ("new speculation" or "prior speculation") and, if not, to propose their own annotation according to the categories "new speculation", "prior speculation", "undetermined speculation", "maybe a speculation but not sure" and 'not a speculation".

Biological speculative sentences have been studied by linguists (Hyland, 1995; Light et al., 2004; Kilicoglu and Bergler, 2008), and some sets of guidelines for annotation have been proposed (Medlock, 2008; Szarvas et al., 2008). However, we did not completely follow these guidelines because we want to categorize sentences (into "new" and "prior" speculation), and because, on contrary to these prior analysis, we are not interested in knowing if a non-explained hypothesis is demonstrated or is supported by some facts (sentence considered as non informative for us), or because we do not consider a sentence as a speculation when the author is being circumspect about some of his statements (for example with the expression "*to our knowledge*"). Indeed, we want especially to annotate ideas and proposals about biological issues, and we consider speculations as a potential source of relevant information for biologists.

4.2. Results of the evaluation

The two randomly selected texts for the evaluation were published respectively in PNAS (first text) and Science (second text) journals and consist of about 400 sentences and 5500 words for the first one and about 200 sentences and 3000 words for the second one. BioExcom annotated exactly 30 sentences in the first text and 29 in the second one[4]. The "correct" annotations were determined on the basis of the set of human annotations. These correct annotations are defined as the most frequent annotations attributed by the evaluators.

The evaluation of speculative sentences' extraction gave the following results (inter-annotator disagreement[5]: 15%):

Precision: 98, 3 %; Recall: 95, 1 %

If we consider also the categorization (prior and new) of speculative sentences, we observe a weak decrease of performance (inter-annotator disagreement: 34%):

Precision: 93, 2 %; Recall: 90, 2 %

Hence, the system finds and categorizes accurately speculative sentences in biological papers. However, these results are preliminary because they should be confirmed at a larger scale. But, we can make some comments on them.

First, it has to be mentioned that despite the annotation guidelines and their careful reading by the evaluators, the task of annotating speculative sentences remains quite subjective or difficult (see inter-annotator disagreement rates). The following sentence was annotated as speculative by one evaluator, although it was an open question:

> *"At present, however, the genes regulating USV function and the development of the cerebellum and the maturation of Purkinje cells are unknown."*

In the same way, the sentence below was wrongly annotated by some evaluators as non speculative, even though it makes a proposal about a biological issue and was correctly annotated by BioExcom thanks to the use of the linguistic markers "*whether"* and "*were not previously established*" by CE processing (see part 3):

> *"Which functional domain of Foxp2 or alternative splicing product of Foxp2 functions in the molecular mechanism of mouse USVs and whether the phenotype of Foxp2-KO mice is due to the loss of function of forkhead domain were not previously established."*

Although the following sentence was clearly a speculation, it has not been annotated by BioExcom, because of the lack of specific linguistic marker (indeed "*should*" can not be a marker strong enough to denote accurately a speculation).

> *"In contrast, heterozygous Foxp2 (R552H)-KI mice, which showed modest impairment of USVs with different USV qualities and which did not exhibit nuclear aggregates, should provide insights into the common molecular mechanisms between the mouse USV and human speech learning and the relationship between the USV and motor neural systems."*

One other limitation is that some linguistic markers are missing in the implementation of BioExcom. This is the case in the following sentence which has not been recognized as a speculative sentence by the system ("in principle" can be a positive clue to disambiguate the indicator "could" but was not yet implemented in BioExcom):

> *"In principle, the act of transcribing Xist could induce structural changes that could alter chromosome wide function (1)."*

We present here one wrong categorization (but correct extraction as a speculative sentence), which has also been performed by BioExcom: the following sentence was categorized as a "new speculation" because of the presence of bibliographic citations.

> *"Foxp2 (R552H) nuclear and/or cytoplasmic aggregates caused ER stress in vitro in cell culture (Fig. 5 E–H), probably because of the polyglutamine region, because similar observations were detected in cells expressing polyQ cytoplasmic aggregates (19)."*

Obviously, we have to perform our evaluation on a larger scale (more texts), but the results of our evaluation are very encouraging.

[4] The automatically annotated corpus is accessible on demand.

[5] One disagreement is recognized when, for all "correct" annotations (human annotations), at least one evaluator is in disagreement with the others.

5. Perspectives

The first perspective is to finalize the system of automatic annotation in order to offer a user-friendly interface. BioExcom will soon be online to be freely used by researchers to annotate their selected papers.

Another project, which is almost completed and will be published very soon, is to enlarge the system so as to index all the words of the database of annotated speculative sentences, as it has previously been done with the EXCOM platform. The user (a researcher in biology) may then look for the presence of a list of keywords in the database of speculative sentences. This will enable him to know all hypothesis or speculations proposed about a biological entity (gene or protein for example) or a biological process, which is very useful for researchers (Light et al., 2004).

Furthermore, speculative sentences have the advantage of being very general, whereas the most powerful and useful text mining systems are often very domain-specific (protein phosphorylation (Yuan et al., 2006) for example), probably, in order to meet biologists' specific needs (Cohen and Hunter, 2008). Nevertheless, because of this specificity, it is obvious that, despite their efficiency to recognize particular patterns in sentences, these systems do not answer entirely the challenge of bridging disjoint literatures. Indeed, most of the time, a knowledge discovery concerns different domains that need to be crossed in a single system in order to establish an unexpected link between two terms (Bekhuis, 2006). To satisfy this requirement, the user of our system may also look for the combination of two lists of terms (Boolean search) in the speculative sentences, in order to find a hypothesis linking these two terms. This link can be either not well demonstrated yet, or unknown, but already discussed or considered in literature from a theoretical point of view. This has been applied to one current discovery, illustrating not only the process by which a prior speculation is becoming a result, but also the possibility to use BioExcom as an aided decision-making system for researchers (unpublished).

It is also important in the future to connect several dictionaries in order to give the possibility to the researcher to enlarge his list of keywords. In the emergent field of "opinion mining", BioExcom would then be able to better highlight papers describing ideas and proposing hypothesis about precise biological issues, which is a tendency in biological literature and a challenge for text mining tools (Rebholz-Schuhmann et al., 2005).

Along these lines, another application of our work that has not been dealt with in the present paper is the use of speculative sentences annotation for discriminating demonstrated biological facts and data from speculative statements. This is an important problem in Text mining classical methods because speculative sentences are very common (30-40 %) in the results, discussions and conclusions sections in biomedical papers (Mercer and Di Marco, 2004). The CE processing, and particularly the use of negative clues for some linguistic markers of speculation, can enable to avoid characterizing, in text mining, a speculation as a definite statement.

References

Alrahabi M, Desclés JP (2008) *Automatic annotation of direct reported speech in Arabic and French, according to semantic map of enunciative modalities*. *In* 6th International Conference on Natural Language Processing, GoTAL, Gothenburg, Sweden, pp 41-51

Bekhuis T (2006) *Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy*. Biomed Digit Libr 3: 2

Bertin M (2008) *Categorizations and annotations of Citation in Research Evaluation*. *In* FLAIRS-21, Miami, Florida, pp 456-461

Blais A, Atanassova I, DesclésJP, Zhang M, Zighem L (2007) *Discourse Automatic Annotation of Texts: an Application to Summarization*. *In* FLAIRS 2007, Special Track "Automatic Annotation and Information Retrieval: New Perspectives", Key West, Florida

Cohen KB, Hunter L (2008) *Getting started in text mining*. PLoS Comput Biol 4: e20

Desclés JP, (2006) *Contextual Exploration Processing for Discourse Automatic Annotations of Texts*. *In* FLAIRS 2006, invited speaker, Melbourne, Florida, pp 281-284

Di Marco C, Kroon FW, Mercer RE (2006) *Using Hedges to Classify Citations in Scientific Articles*. *In* Computing Attitude and Affect in Text: Theory and Applications, Vol 20, pp 247-263

Djioua B, Flores JG, Blais A, Desclés JP, Guibert G, Jackiewicz A, Le Priol F, Nait-Baha L, Sauzay B (2006) *EXCOM: an automatic annotation engine for semantic information*. *In* FLAIRS 2006, Melbourne, Florida, 11-13 mai, pp 285-290

Hunter L, Cohen KB (2006) *Biomedical language processing: what's beyond PubMed?* Mol Cell 21: 589-594

Hyland K (1995) *The author in the text: Hedging Scientific Writing.* Hong Kong papers in linguistics and language teaching 18: 33-42

Kilicoglu H, Bergler S (2008) *Recognizing speculative language in biomedical research articles: a linguistically motivated perspective.* BMC Bioinformatics 9 Suppl 11: S10

Le Priol F (1999) *A data processing sequence to extract terms and semantics relations between terms. In* HCP'99 (Human Centered Processes-10th Mini EURO Conference), Vol 241-8

Light M, Qiu XY, Srinivasan P (2004) *The Language of Bioscience: Facts, Speculations, and Statements in Between. In* HLT-NAACL, ed, Workshop On Linking Biological Literature Ontologies And Databases, pp 17-24

Medlock B (2008) *Exploring hedge identification in biomedical literature*. J Biomed Inform 41: 636-654

Mercer RE, Di Marco C, (2004) *A Design Methodology For A Biomedical Literature Indexing Tool Using The Rhetoric Of Science. In* LT-NAACL Workshop: Biolink 2004, Linking Biological Literature Ontologies And Databases; 2004.

Rebholz-Schuhmann D, Kirsch H, Couto F (2005) *Facts from text--is text mining ready to deliver?* PLoS Biol 3: e65

Rzhetsky A. SM, Gerstein M. (2008) *Seeking a new biology through text mining.* Cell 134: 9-13

Szarvas G, Vincze V, Farkas R, Csirik J (2008) *The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In* BioNLP ACL-2008 workshop,

Teufel S (1998) *Meta-Discourse Markers And Problem-Structuring In Scientific Articles. In* Workshop On Discourse Relations And Discourse Markers,

Wilbur WJ, Rzhetsky A, Shatkay H (2006) *New directions in biomedical text annotation: definitions, guidelines and corpus construction.* BMC Bioinformatics 7: 356

Yuan X, Hu ZZ, Wu HT, Torii M, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH (2006) *An online literature mining tool for protein phosphorylation.* Bioinformatics 22: 1668-1669