Discursive Mining Viewpoints in Building Multi-Document Synthesized Sheets

Olfa Makkaoui*, Jean-Pierre Desclés*,
Marc Bertin*, Christophe Jouis, Jean-Gabriel Ganascia****
∗LALIC-STIH, Université Paris-Sorbonne, 28, rue Serpente - 75006 Paris, France
∗∗LIP6, Université Pierre et Marie Curie and CNRS, 4, place Jussieu - 75005 Paris, France

Abstract

Multi-documents sheets are viewed as semantically structured representations of textual documents. The automatic construction of these sheets is based on the automatic annotation of textual documents according to a set of discursive categories called discursive mining viewpoints. The automatic annotation of a text is performed using the Contextual Exploration processing. It is a linguistic and computational method implemented in the EXCOM2 platform that allows the annotation of segments (which can be a title, a paragraph, a sentence or a clause) according to a given discursive mining viewpoint.

Introduction

Multi-documents sheets are viewed as semantically structured representations of textual documents. The automatic construction of these sheets is based on the automatic annotation of textual documents according to a set of discursive categories called discursive mining viewpoints such as : extracting definitions given by different scientists about a same notion from a corpus of texts relative to a specific area of knowledge; extracting new assumptions (compared to old assumptions) about a previously studied subject; discovering new and recent results; identifying the most effective and operational methods used in experiments of a scientific field; identifying the plausibility of an hypothesis; identifying citations of a specific author (in order to better answer to the question" how, and why, an other author's works are are cited by other authors ?"); identifying, in all publications of an author, the different quotations of direct and indirect reported speech of other authors (to answer to the question :"how an author reports the quotations of other authors ?"). The automatic annotation of a text is based on a linguistic and computational technique : The Contextual Exploration processing (designated henceforth by CE) for the semantic analysis of textual documents. This linguistic technique is executed by the automatic annotation engine EXCOM2. The linguistic technique CE and the EXCOM2 engine do not require prior morphological or syntactic analysis.

Why synthesized sheets ?

Our objective is to present a general technique that enables to automatically extract information and build structured knowledge relating to a same research field. This technique requires an automatic process that annotates, exctracts and storages the annotated sentences in structured sheets according to different discursive mining viewpoints. By consulting synthetized sheets, users can extract most relevant information found in a corpus and manage knowledge of a studied subject. Thus, in order to obtain this information, a particular user can insert a set of new requests using specific term (or a set of synonymous terms or named entities) related to his research field. This enables him to filter among the already built synthetized sheets a specific knowledge which are presented as a set of structured annotated related to the searched terms.

The profiles of users of synthesized sheets are multiple. An user can be a researcher or a professional of information research with targeted needs (for instance : to know the recent assumptions about Alzheimer's disease with a short review of articles published in the last six months). Users who interested in constructing automatically synthetized sheets are : (i) A researcher or a professional of information research with targeted needs (for example, extracting recent assumptions concerning the Alzheimer's disease from a short review of articles published in the last six months). (ii) A student who aims to collect information in order to prepare a presentation or a scientific paper that describes the recent evolution of a scientific area. (iii) An engineer who is building a domain ontology from an analysis of different textual documents. (iv) An appraiser who aims to identify rapidly new results and innovations obtained in a laboratory in order to decide whether he allocates a financial assistance to its research. Many approaches seek to manage knowledge extracted from several publications (not only full texts but also abstracts) or from a big textual document (such as an academic thesis, a book, a technical report…). Some approaches use keywords or are based on statistical criteria (based for example on the identification of the most frequent linguistic expressions), or on machine learning approaches. Other methods, called "linguistic methods", require morphological and syntactical analysis, using automatic linguistic platforms like GATE. These linguistic methods have to deal with problems linked to different morphological and syntactical ambiguities that appear in some sentences. When a user have to explore many textual documents, reading abstracts requires a lot of time, and is not enough for users who want to stay up to date with all the developments in the fields they are intrested in. Indeed, without an automatic process that is able to mine a large amount of information, the user must search in large databases of abstracts and often turn back to the original full texts in order to find out what really interests him. Furthermore, a user needs to contextualize information obtained in extracted sentences (to distinguish between new and prior results, new hypothesis and new results. . .), to interpret and appraise them. It should also be noted that an information with a very low frequency can be informative for a domain expert but not considered as important in some textual summaries. Indeed, the domain expert needs to discover new knowledge by crossing information extracted from different documents or research area (Bekhuis 2006) where a hidden information has low frequency in the text. Our objective is to constuct an easily reachable tool that builds synthetized sheets categorized according to discursive mining viewpoints. A user can for example ask the following questions:

- “What is the method used by the author X ?”;
- “What are the results obtained by means of the method Y ?”;
- “What are the specific hypotheses claimed by the author X ?”;
- “What are the new hypotheses (and not the old hypotheses) and the new results about the studied object Z ?”;
- “What are the deep features underlined by the author X about the studied object Z ?”;
- “What is the restrictions (or no restrictions) used by the author X when he has cited the author X’ in his article ?”;
- “ What is the quotation of X’ and how it is reported by the author X ?”. . .

The aim of the synthetized sheets building is : to construct an effective automatic tool for textual documents semantic mining. These documents are generally scientific publications relative to a same field of knowledge. The automatic synthetized sheets construction is based on the classification of the extracted annotated sentences according to their discursive and semantic categories (such as definition, hypothesis, result, methods, bibliographic citations (Bertin 2011)). Synthetized sheets are usefull since they enable users to access rapidly to a categorized information, to cross extracte information from different textual documents, to allow discovering new knowledge and to collect obtained information in a same synthetized sheet. In order to obtain a structured summary of many documents a user who has already collected a set of textual documents, uses this tool with the following functionalities : (i) The automatic semantic annotation of texts according to different discursive categories or discursive mining viewpoints; (ii) The storage of annotated texts in synthetized sheets; (iii) The navigation in synthetized sheets obtained from supplementary information given by users (such as terms related to a studied subject, an author name, a specific method or a class of equivalent named entities).

Contextual Exploration and automatic discursive annotation engine EXCOM 2

The CE is a linguistic and computational technique (Desclés 1997) that is directly oriented to semantic text mining applications. It has already been presented in many publications (Jouis 1993; Minel and Desclés 2000; Desclés 2006; Jean-Pierre and Florence 2009). The construction of the CE rules is performed by the EXCOM2 annotation engine (Alrahabi 2010) and depends on the linguistic resources of a given discursive mining viewpoint (speculation for the BioExcom system). The constructed rule, written in the XML format, are then used by the EXCOM engine to annotate textual segments. First, texts are segmented into sentences using a list of typographical signs. This step takes as input text files and returns segmented texts in the XML format. Then, the CE rules are applied to annotate textual segments by adding meta-data information to sentences of the segmented files when the annotation rules are confirmed. The important features of EXCOM2 are:

(i) It does not require previous morphological and syntactical analysis avoiding possible ambiguities (morphological or syntactical) in a sentence;

(ii) It is independent of specific scientific domains (agronomy, geology, biology, physics, mathematics. . .) but also human sciences as linguistics, psychology, sociology. . .). It performs with only linguistic markers associated to discursive categories used for a semantic textual mining (for instance, the linguistic markers which are used to identify a new hypothesis in a text are the same in different domains).

EXCOM2 is a rule-based system that uses a set of rules triggered inside of an analyzed text, by a recognition of occurrences of linguistic markers associated to discursive mining viewpoints (named ”identificators” of discursive mining viewpoint) and other linguistic expressions (”complementary indices”) by means of CE rules in the context of these occurrences. The different conceptswhich are more or less specific to a general discursive minig viewpoints are organized in a ”semantic map”, or a linguistic ontology. To each node in the semantic map of a given discursive mining viewpoint is associated a more specific concept. For example, ”new hypothesis” or ”prior hypothesis” can be seen in the text as more specific linguistic indicators, which are the linguistic markers of the concept” hypothesis”. Since indicators of a discursive mining viewpoint are often ambiguous, to avoid an important noise, it is needful to identify other specific linguistic clues in a text segment (e.g. a sentence). This segment is the linguistic context of an indicator. A set of CE rules is associated to each indicator or to a class of equivalent indicators rules. A CE rule seeks for additional linguistic markers (”clues”) in the context of an indicator in order to annotate the textual segment where the indicator has one occurrence. Indicators, rules and clues are instances of the concept set in the semantic map. We give as example the summarization semantic map presented in 2.

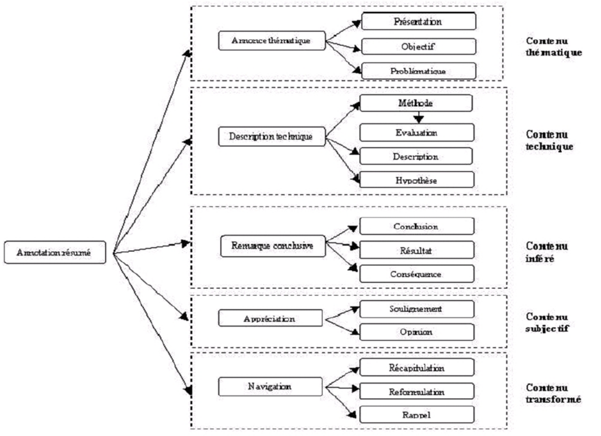


Figure 1: Textual summarization semantic map (Blais 2008)

Basically, the annotation of a text by EXCOM2 requires the following steps: 1) Automatic segmentation of texts into sections, paragraphs and textual segments; 2) Search for indicators in the segment for a selected discursive mining viewpoint by a finite automata and then call and execution of the associated CE rule which are triggered by the identification of an indicator in the textual segment. 3) Search for additional linguistic markers contained in the rule. This search is performed in the sentence research space (at the right or/and the left of the indicator or even inside the indicator) according to the rule 4) Semantic annotation of the segment if all the rules conditions are filled.

The automatic annotation process is already implemented and evaluated in several discursive mining viewpoints :

The automatic annotation process has already been implemented and evaluated with different resources relative to different discursive mining viewpoints :

- The identification of relations between concepts in a text in order to populate a domain ontology (Jouis 1993);
- The identification of bibliographic citations with an indication of the judgment (agreement or not) and appraisals of the author of the article, according to the discursive mining viewpoint -"How someone is cited ?"- (Bertin 2008; 2011);
- The identification of reported discourse with the specification of the contextual conditions of the identified quotation, according to the discursive mining viewpoint How someone is quoted by a direct and indirect reported speech in french and arabic texts (Alrahabi 2010);
- The identification of causal relations, definitions in texts;
- The automatic segmentation of texts (In paragraphs, sentences, textual segments) according to the discursive mining viewpoints : how the punctuation analysis can be used to segment a text (Mourad 2002);
- The automatic summarization in French and English (Berri et al. 1996; Blais, Desclés, and Djioua 2006; Blais et al. 2007; Blais 2008).

Semantic relations between authors

We identify the semantic relation between authors in textual segments containing indexed references and use an automated semantic annotation platform, EXCOM2, to annotate our corpora. The main categories are organised in a semantic map as in Bertin 2008 and it is fully operational in this implementation. The categories in this semantic map will be used to classify the scientific texts. This approach allows the annotation of the segments containing indexed references. The corpora contains mainly scientific texts and articles available from journals [1]. In this work, we explore 288 scientific published papers ; segmented corpus into 77,000 sentences. We automatically annotate semantically over 500 sentences. We will here set few examples from our French corpus. Examples correspond to the definition discursive mining viewpoint.

- *"Aujourd'hui, l'habitat de la girafe au Niger tel que defini par Kawa (2000) comprend : . . . "*
- *"La prédiction spatiale de la distribution des espèces reflété alors le concept de niche écologique définie par Hutchinson [25]"*
- *"Nous avons retenu comme définition de la cécité binoculaire [2] : tout enfant présentant une acuité visuelle inférieure a 3/60 du meilleur oeil avec la correction portée."*

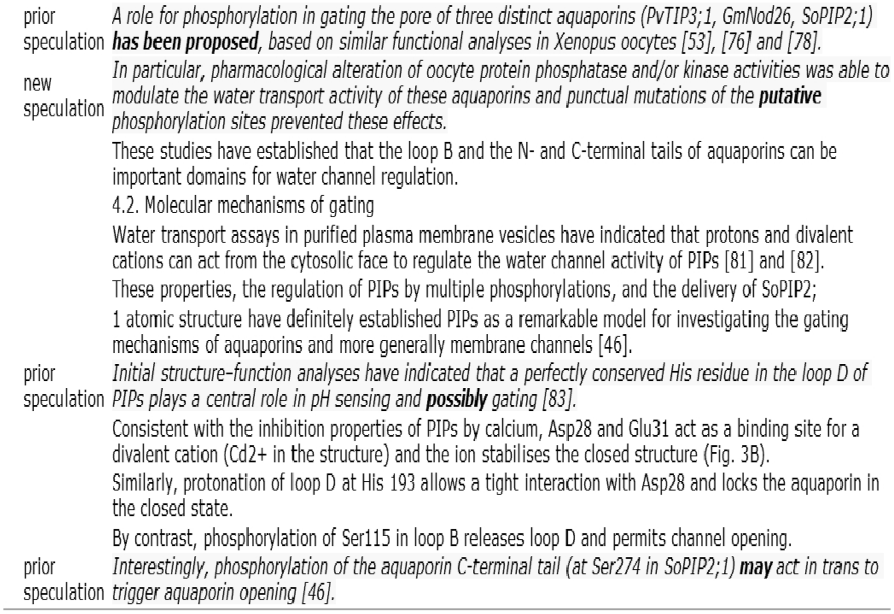
An important point of this approach is that it is now possible to design from textual segments a representation from problematic around the area of health, agriculture or drought.

Automatic identification of speculation (plausible hypothesis) in biomedical papers : Building synthetic sheets and evaluation of the automatic process Speculation in biomedical papers

In order to implement and to evaluate our method for building synthetized sheets, we have chosen to study a particular discursive mining viewpoint, the "speculation identification" in biological papers. A "speculation" is a plausible proposal, not observed or not deducted directly and explicitly presented as not certain in the text (See the figure 3). Biologists are particularly interested in knowing all the speculation expressed about a biological entity or a specific topic since speculation may suggest other ways of looking at a problem and to guide a research program to new experiments (Blagosklonny and Pardee 2002; Bray 2001; Light, Qiu, and Srinivasan 2004; Medlock and Briscoe 2007). This work about synthesized sheets in the biological field is presented and discussed in several international publications (Desclés, Alrahabi, and Desclés 2009; 2011).

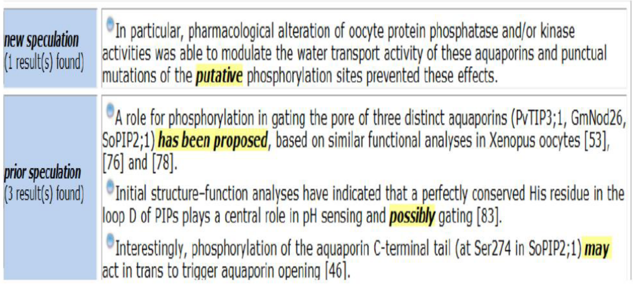
Linguistic ressources of the speculation discursive mining viewpoint (a semantic map with linguistic indicators, associated CE rules and linguistic clues) are developed based

[1] John Libbey Eurotext edition allow us to create a corpus and to annotate their scientific journals.



prior speculation	*A role for phosphorylation in gating the pore of three distinct aquaporins (PvTIP3;1, GmNod26, SoPIP2;1)* ***has been proposed****, based on similar functional analyses in Xenopus oocytes [53], [76] and [78].*
new speculation	*In particular, pharmacological alteration of oocyte protein phosphatase and/or kinase activities was able to modulate the water transport activity of these aquaporins and punctual mutations of the* ***putative*** *phosphorylation sites prevented these effects.*
	These studies have established that the loop B and the N- and C-terminal tails of aquaporins can be important domains for water channel regulation.
	4.2. Molecular mechanisms of gating
	Water transport assays in purified plasma membrane vesicles have indicated that protons and divalent cations can act from the cytosolic face to regulate the water channel activity of PIPs [81] and [82].
	These properties, the regulation of PIPs by multiple phosphorylations, and the delivery of SoPIP2;
	1 atomic structure have definitely established PIPs as a remarkable model for investigating the gating mechanisms of aquaporins and more generally membrane channels [46].
prior speculation	*Initial structure–function analyses have indicated that a perfectly conserved His residue in the loop D of PIPs plays a central role in pH sensing and* ***possibly*** *gating [83].*
	Consistent with the inhibition properties of PIPs by calcium, Asp28 and Glu31 act as a binding site for a divalent cation (Cd2+ in the structure) and the ion stabilises the closed structure (Fig. 3B).
	Similarly, protonation of loop D at His 193 allows a tight interaction with Asp28 and locks the aquaporin in the closed state.
	By contrast, phosphorylation of Ser115 in loop B releases loop D and permits channel opening.
prior speculation	*Interestingly, phosphorylation of the aquaporin C-terminal tail (at Ser274 in SoPIP2;1)* ***may*** *act in trans to trigger aquaporin opening [46].*

Figure 2: Example of an annoted text according to the speculation data mining viewpoint



new speculation (1 result(s) found)	In particular, pharmacological alteration of oocyte protein phosphatase and/or kinase activities was able to modulate the water transport activity of these aquaporins and punctual mutations of the ***putative*** phosphorylation sites prevented these effects.
prior speculation (3 result(s) found)	A role for phosphorylation in gating the pore of three distinct aquaporins (PvTIP3;1, GmNod26, SoPIP2;1) ***has been proposed***, based on similar functional analyses in Xenopus oocytes [53], [76] and [78]. Initial structure-function analyses have indicated that a perfectly conserved His residue in the loop D of PIPs plays a central role in pH sensing and ***possibly*** gating [83]. Interestingly, phosphorylation of the aquaporin C-terminal tail (at Ser274 in SoPIP2;1) ***may*** act in trans to trigger aquaporin opening [46].

Figure 3: Visualization of asynthetized sheet of the speculation data mining viewpoint: Annotated sentences are categorized into new and prior speculation

on a small representative corpus analysis, in order to extract pertinent linguistic markers and then to formulate the CE rules. The speculation discursive mining viewpoint consists of twelve indicator classes (same semantic or grammatical categories) and thirty rules. Thus, in order to annotate automatically textual segment the BioExcom system has been constituted; it annotates textual segment with the EXCOM2 engine. To identify speculative sentences, BioExcom uses a set of indicators such as ”suggest”, ” may” or contiguous patterns as ”we hypothesize that”. For example, the presence of the indicator ”We hypothesize that” in sentence (1) allows its annotation as ”speculation”.

(1) ”We hypothesize that a mutation of the hGR glucocorticoid-binding domain is the cause of cortisol resistance”.

However, indicators are often ambiguous and their presence in a sentence does not automatically implicate that it is possible to annotate it according to a given discursive mining viewpoint. To remove ambiguities and noisy annotations, it is useful to search for some additional linguistic markers (clues) in the context of an indicator in order to confirm (or to invalidate) the semantic decision (to annotate or not a textual segment). These clues are not always very close to the indicator in the textual segment. For example, sentences (2) and (3) have the same indicator ”is unclear”. However, the sentence (2) deals with a lack of knowledge while the sentence (3) is used to express a speculation due to the presence of the additional marker ”whether”:

(2) ” The precise role of such ligninolytic enzymes is unclear because none of them is able to delignify intact lignocellulose in vitro.”

(3) ”Also, it is unclear whether the measures used, such as high blood pressure, succeed in capturing the underlying biological processes, or are outcomes associated with physiological breakdown.”

Evaluations

To evaluate the BioExcom performance, we use a large corpus of manually annotated biomedical papers : The Bioscope corpus (Vincze et al. 2008). The evaluated corpus contains 1273 abstracts and 9 full texts annotated manually and independant from linguistic ressources of the BioExcom system. Following the BioExcom criteria used to identify speculation by BioExcom (Desclés, Alrahabi, and Desclés

2009), we compare the manual annotations of BioScope with annotations obtained automatically by BioExcom. The comparison showed some annotation errors in BioScope. The application of BioExcom on a corrected corpus has an F-score of 90.1% (82.7% recall and precision of 99.1%) for the automatic detection of speculation (Desclés, Alrahabi, and Desclés 2009; 2011). It should also be noted that the BioExcom system (linguistic ressources executed with the EXCOM2 engine) has, in particular, annotated very rapidly large biomedical papers with a short processing time. Thus, the positive evaluation results of the BioExcom for the automatic annotation of "speculation" discursive mining viewpoint allows in a second phase to extract automatically annotated sentences and to index them in order to obtain synthesized sheets. These latter can be also built by crossing between annotations and specific information filled by the user (such named entities). Once synthetized sheets are built, it is possible to return to the original text (from which annotated sentences were extracted). This enables the user to verify the relevance of the extracted information and eventually to read the analyzed paper.

Conclusion

In this paper we focused on the usefulness of the synthesized sheets construction from different discursive mining viewpoints which are organized in a semantic maps. Synthetized sheets allow user to extract information from large corpus of textual documents (scientific publications, doctoral dissertations, books, technical reports…). These synthetized sheets are usually realized maunally by researchers who have to collect information about a specific subject : to analyse definitions of theoretical terms by systematic comparisons of different definitions proposed by several writers; to know recent and new results about a studied object (for instance : a molecule, a protein in biology; a controversial concept in social sciences; a specific writer…). An automatic tool that is able to automatically constitute synthesized sheets is usefull for different users (researchers, specialists of information, students…). The extracted information are stored in synthetized sheets which enables users to extract specific information from a significant number of documents according to discursive mining viewpoints. This process is more rapid and efficient than mining all the document sentences using a set of keywords. For example, the extraction of new hypothesis about a diatom by crossing information from a set of biomedical papers can reveal new properties about this diatom which are not found using simple keywords. We have shown in this paper how to build automatically synthesized sheets from automatic semantic annotations of corpora (scientific publications). With different discursive mining viewpoints, it is possible to constitute linguistic ressources. First, linguistic ressources are given by a semantic map of a specific discursive mining viewpoint. Then indicators classes and CE rules (with precise linguistic cues) associated to each concept (or more specific discursive mining viewpoint) are organized in the semantic map. The evaluation of discursive mining viewpoints (such as speculation) demonstrated that the CE annotation process can give good results. This discursive mining viewpoints was evalutated on a large corpus that was annotated manually for speculation. We conclude from the prevoius experiment that the annotation by EXCOM2 has very good performance in processing time and volumes of documents as well as the accuracy of annotation system. It allows an effective building of multi-document synthesized sheets. Our future work is to develop a set of general discursive mining viewpoints that is not dependant on a specific field but that covers also other research areas like humain and social science.

References

Alrahabi, M. 2010. *EXCOM-2 : plate-forme d'annotation automatique de catégories sémantiques: Applications à la catégorisation des citations en français et en arabe.* Ph.D. Dissertation, Université Paris-Sorbonne.

Atanassova, I. 2012. *Exploitation informatique des annotations sémantiques automatiques d'Excom pour la recherche d'informations et la navigation.* Ph.D. Dissertation, Université Paris-Sorbonne.

Bekhuis, T. 2006. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries* 3(1):2.

Berri, J.; Cartier, E.; Desclès, J.-P.; Jackiewicz, A.; and Minel, J.-L. 1996. *SAFIR, système automatique de filtrage de textes.* Ph.D. Dissertation, Université Paris-Sorbonne.

Bertin, M. 2008. Categorizations and Annotations of Citation in Research Evaluation. In *FLAIRS Conference*, 456–461.

Bertin, M. 2011. *Bibliosémantique : une technique linguistique et informatique par exploration contextuelle.* Ph.D. Dissertation, Université Paris-Sorbonne.

Blagosklonny, M. V., and Pardee, A. B. 2002. Conceptual biology: unearthing the gems. *Nature* 416(6879):373–373.

Blais, A.; Atanassova, I.; Desclés, J.-P.; Zhang, M.; and Zighem, L. 2007. Discourse automatic annotation of texts: An application to summarization. In *FLAIRS Conference*, 350–355.

Blais, A.; Desclés, J.; and Djioua, B. 2006. Le résumé automatique dans la plate-forme excom. In *Digital Humanities*.

Blais, A. 2008. *Résumé automatique de textes scientifiques et construction de fiches de synthèse catégoriesées : approche linguistique par annotations sémantiques et réalisation informatique.* Ph.D. Dissertation, Université Paris-Sorbonne.

Bray, D. 2001. Reasoning for results. *Nature* 412(6850):863–863.

Desclés, J.; Alrahabi, M.; and Desclés, J.-P. 2009. BioExcom: Automatic Annotation and categorization of speculative sentences in biological literature by a Contextual Exploration processing. In *Proceedings of the 4th Language & Technology Conference*.

Desclés, J.; Alrahabi, M.; and Desclés, J.-P. 2011. BioExcom: Detection and categorization of speculative sentences in biomedical literature. In *Human Language Technology.*

Challenges for Computer Science and Linguistics. Springer. 478–489.

Desclés, J.-P. 1997. Systèmes d'exploration contextuelle. *Co-texte et calcul du sens* 215–232.

Desclés, J.-P. 2006. Contextual exploration processing for discourse and automatic annotations of texts. In *FLAIRS Conference*, 281–284.

Djioua, B.; Flores, J. J. G.; Blais, A.; Desclés, J.-P.; Guibert, G.; Jackiewicz, A.; Le Priol, F.; Nait-Baha, L.; and Sauzay, B. 2006. Excom: An automatic annotation engine for semantic information. In *FLAIRS Conference*, 285–290.

Djioua, B.; Desclés, J.; and Alrahabi, M. 2012. Next generation search engines: Advanced models for information retrieval, chapter searching and mining with semantic categories. *IGI Global* 1:115–137.

Jean-Pierre, D., and Florence, L. P. 2009. *Annotations automatiques et recherche d'information (Traité Cognition et Traitement de l'Information-IC2)*. Hermes Science Publications. 171–192.

Jouis, C. 1993. *Contributions à la conceptualisation et à la Modélisation des connaissances à partir d'une analyse linguistique de textes: réalisation d'un prototype*. Ph.D. Dissertation, Université Paris-Sorbonne.

Light, M.; Qiu, X. Y.; and Srinivasan, P. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, 17–24.

Medlock, B., and Briscoe, T. 2007. Weakly supervised learning for hedge classification in scientific literature. In *ACL*, volume 2007, 992–999.

Minel, J.-L., and Desclés, J.-P. 2000. *Résumé Automatique et Filtrage des textes*. Hermes Science Publications.

Mourad, G. 2002. *La segmentation de textes par exploration contextuelle automatique, présentation du module segatex*. Ph.D. Dissertation, Université Paris-Sorbonne.

Vincze, V.; Szarvas, G.; Farkas, R.; Móra, G.; and Csirik, J. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics* 9(Suppl 11):S9.