Towards automatic thematic sheets based on discursive categories in biomedical literature

Julien Desclés
LaLIC Université Paris-Sorbonne
Maison de la Recherche
28 rue Serpente, 75006 Paris
julien.descles@gmail.com

Olfa Makkaoui
LaLIC Université Paris-Sorbonne
Maison de la Recherche
28 rue Serpente, 75006 Paris
olfa_makkaoui@yahoo.fr

Jean-Pierre Desclés
LaLIC Université Paris-Sorbonne
Maison de la Recherche
28 rue Serpente, 75006 Paris
tel : (33) 1 53 10 58 25
jean-pierre.descles@paris.sorbonne.fr

ABSTRACT

Biological papers contain a huge amount of results and ideas that are difficult to manage. Researchers are not only interested in finding relevant information but they also need to know how the authors of papers get the results. Thus, it is interesting to identify the methods used and to distinguish for example between speculations, observations and deductions. Biologists need also to distinguish between new and prior information, especially to identify the real new output of a study. In order to respond to these needs, we propose a linguistic model based on the discursive categories. This model aims to develop the BioExcom tool for the automatic production of thematic sheets using the Contextual Exploration processing. BioExcom is already able to detect speculative sentences and to categorize them into new and prior speculation. The other categories of the model will be developed using the proposed linguistic markers.

Categories and Subject Descriptors

D.3.3 .3.1 [**Information storage and retrieval**]: Content analysis indexing – *Abstracting methods, linguistic processing;* I.2.7 [**Artificial intelligence**]: Natural language processing – *Discourse, text analysis* I.7.5 [**Document and text processing**]: Document capture – *Document analysis;* J.3 [**Life and medical science**]: *Biology and genetics, Medical information systems.*

General Terms

Language, Documentation.

Keywords

Discourse analysis, Biology, Automatic annotation, Categorization, Contextual exploration, Summarization, Information extraction

WIMS'11, May 25-27, 2011 Sogndal, Norway
Copyright © 2011 ACM 978-1-4503-0148-0/11/05... $10.00

1. INTRODUCTION

Dealing with a large amount of information in biological scientific literature provided by search engines such as PubMed/MEDLINE, which references more than 19 million biological papers, requires the use of Information Extraction techniques. Several applications use machine learning techniques, rule-based systems or also a combination of these (see [1] [2] [3]). The main purpose of these applications can be to find the appropriate documents, to extract specific information or to automatically summarize papers.

Most of the more useful systems (see [4]) focus on very specific and particular tasks. RLIMS-P, for example, recognizes sentences dealing with proteins phosphorylation [5], or STRING-IE system extracts gene and proteins regulation networks [6]. Despite all these tools, it is very difficult for a researcher to keep posted on all the developments in the fields which could interest him. Indeed, one challenge in text mining and knowledge discovery is to bridge disjoint literatures (different domains). In addition, a researcher does not want only to extract information but also to decide which publication deserves to be read and to know how the authors get their new results.

For example, the results of a search for information linked to particular gene name using the previous mentioned Information Extraction systems do not distinguish between hypothesis, new and prior results. Thus, [7] underlined the importance of the kind of information in biological papers by explaining that:

"*The fact that a gene is mentioned in the text and the text states, for example, that a gene is regulated by another gene, does not necessarily imply that the information is reliable or useful*"

In this paper, we describe the BioExcom project which uses the Contextual Exploration processing to annotate automatically biological papers. We first propose a linguistic model based on discursive and semantic categories and short lists of markers specific to each category. We then detail the current progress of BioExcom dealing with the automatic detection and categorization of speculation.

2. RELATED WORK

[8] focused on the automatic text summarization of scientific papers by using rhetorical categories. They proposed a model based on seven categories (Background, Other, Own, Aim, Textual, Contrast and Basis). These categories are also used by [9] who improved the model by proposing their own seven classes (Background, Problem, Outline, Textual, Own, Connexion, and Difference). [10] performed summarization and thematic sheets[1] of scientific papers by focusing on the automatic annotation of texts according to discursive categories based on texts linguistic analysis.

[11] provided a theoretical model to categorize scientific papers organized according to six linked classes: Meta-information, Positioning, Method, Result, Interpretation and Conclusion. [12] studied the ABCDE (Annotation, Background, Contribution, Discussion and Entities) rhetorical structure of scientific papers and identified seven types of epistemic segments: Fact, Hypothesis, Implication, Goal, Method, Result and Problem.

Some previous works have focused on the certainty level to categorize sentences. In [13] certainty in newspaper articles is annotated according to four dimensions: Degree of certainty, Focus, Perspective and Time. Also, the model proposed by [14] seeks to classify sentences annotated as biological event according to three dimensions: Knowledge Type (demonstrative, deductive, sensory and speculative), Certainty Level (absolute, high, moderate and low) and Point of view (writer and other).

3. METHODOLOGY

We accomplished, as part of an interdisciplinary collaboration between biologists, linguists and computer scientists, a linguistic analysis on about thirty papers (from very different biomedical journals) to determine the semantic and discursive categories[2] that we are going to develop in the BioExcom system. A set of linguistic marker was extracted from the analysed corpus according to each semantic category (see Table1). Based on this analysis and helpful discussions between biologists and linguists, we also used synonym dictionaries such as "synonym"[3] to enlarge, in some cases, the markers sets.

4. DISCURSIVE AND SEMANTIC CATEGORIES

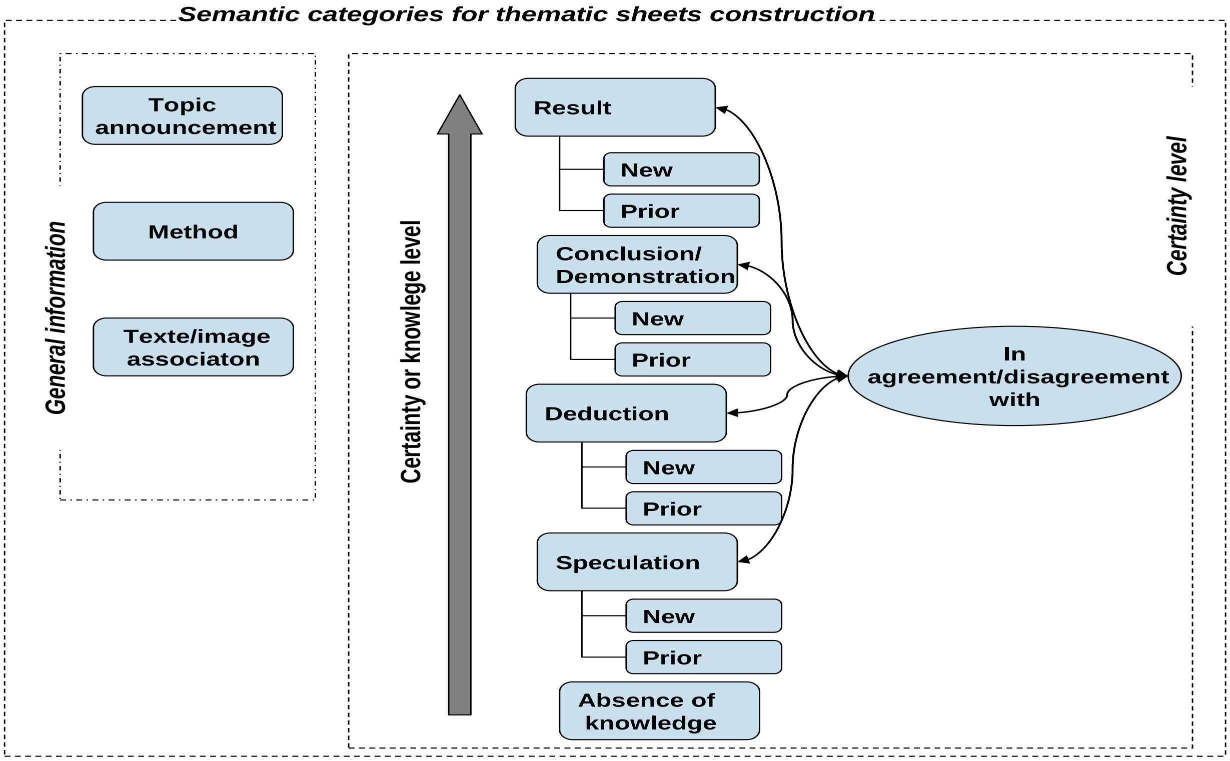

Figure 1. Semantic categories used for thematic sheets construction

[1] A thematic sheet is a thematic summary that classifies and regroups information according to their semantic categories.

[2] A discursive category is linked to the nature of the scientific discourse and is independent of the text domain (definition, conclusion, speculation...)

[3] http://www.synonym.com/

Figure 1 illustrates the proposed linguistic model used to construct automatically thematic sheets. We distinguish between two types of information:

- *General information*: The presented categories (Topic announcement, Method, Text/image association) enable the reader to have a general idea about the paper.
- *Certainty or Knowledge level based categories*: These categories are constructed and ordered according to the certainty/knowledge level that they express (the highest level is the *Result* category).

The main linguistic markers collected in this study are presented in Table 1.

4.1 General information

4.1.1 Topic announcement

The *Topic announcement* category deals with the presentation of examined questions. It provides to the user an idea about the content of the paper which can be more precise than the information presented in the paper title.

Examples:

(1) "*To understand the mechanism by which PRL regulates the biphasic expression of IRF-1, we cloned the rat IRF-1 gene and functionally characterized the IRF-1 promoter."*

(2) "*The major goal of this study was to determine whether the mycobacterial cell wall component mannose-capped lipoarabinomannan (ManLAM) of Mycobacterium tuberculosis (M. tuberculosis) could activate transcription of HIV-1 in T cells with the use of an in vitro cell culture system."*

Linguistic markers used to express topic announcement in biological papers are verbs (such *as to address, to analyze, to attempt, deal with, to describe, to evaluate, to examine,*) or nouns (*aim, idea, intention, objective, purpose*).

4.1.2 Method

The Method category introduces the techniques used to develop the results presented in the paper. This category can be useful for the experimental or theoretical framework of biologists.

Examples:

(3) "The size of the very fine granules or rugosity sometimes observed on the valve surfaces was estimated using atomic force microscopy."

(4) "Solid-state 29Si nuclear magnetic resonance spectroscopy techniques confirmed the amorphous nature of the frustule and were used to estimate the coordination state of the silicone [36 and 37]."

Sentences dealing with methods are marked by verbs such as *use, monitor, measure, perform, assay*. We also notice that the most part of the verbs used in this category are in the passive voice. Method category sentences can be marked by nouns such as: *tool, method* and *technology.*

4.1.3 Text-image association

The identification of the non textual elements (figure, pictures, tables…) in the biological papers is crucial for the thematic sheets construction because a lot of results, particularly in genomic and post-genomic articles, are listed in images.

In addition, many researchers prefer only to look at figures in the articles in order to have a more definite idea about their content (direct access to data) instead of reading the full text. It is also important to extract the sentences making a comment on the figures inside the text. This gives access to very important information which often consists in results. The linguistic markers for this category can be for example: *Fig x, Tab x, Doc x, see the figure above.*

4.2 Certainty/Knowledge level based categories

4.1.3 Result

The Result category illustrates author's observation and is considered very important for researchers. Biologists are especially interested in this type of information that can be very well supplemented by Text-image association category (see above).

Examples:

(5) "We describe the isolation and characterization of multiple cDNAs encoding mouse Oct2 from a mature B-cell line and we show that a variety of isoforms of this transcription factor is generated from a single gene by an alternative splicing mechanism."

(6) "Previously, transcription of the c-fos gene has been reported to be transactivated by the viral transcription factor, Tax1."

Result linguistic markers include verbs like *report, reveal, show, discover, find,* and nouns such as *finding, outcome* and *result.* Obviously, all the results will not be detected by this approach, but *result* markers are not dependent of specific domains (e.g. proteins phosphorylation) and will give access to the most relevant results (highlighted by authors).

4.1.4 Conclusion/Demonstration

Conclusion and demonstration category highlights relevant information and use persuasive techniques to achieve a precise goal which is mostly linked to the topic announcement. We consider that a conclusion/demonstration has a level of certainty less important than a result because it mostly assumes an interpretation of result.

Examples:

(7) "*Our study demonstrated that biological AFM with a live bioprobe can be successfully applied to carry out in situ characterization of cell adhesion to different surfaces."*

(8) "In conclusion, AIDS patients with hypercortisolism and clinical features of peripheral resistance to glucocorticoids are characterized by abnormal glucocorticoid receptors on lymphocytes."

GENERAL INFORMATION		**CERTAINTY/KNOWLEDGE LEVEL BASED CATEGORIES**			
Topic announcement	**Method**	**Result**	**Conclusion/ Demonstration**	**Deduction**	**Absence of Knowledge**
aim	use	confirm	prove	infer	be not known
goal	measure	report	demonstrate	indicate	remains unknown
idea	monitor	identify	conclude	deduce	be not clear
intention	method	detection	state	implicate	no evidence for
purpose	tool	outcome	summary	deduction	further studies are necessary
to address	technique	result	assert	signal	be unknown
to analyse	perform	reveal	demonstration	consequence	be an open question
deal with	assay	show	report		
to describe		discover	attest		
to discuss		find			
to examine		finding			
focus on					
investigate					
present					
to understand					
objective					
to evaluate					

Table 1. Main linguistic markers (indicators) of some semantic categories

Sentences belonging to conclusion can be denoted for instance by *assert, attest, certify, conclude, demonstrate* as verbs and *conclusion, summary* as nouns.

4.1.5 Deduction

The Deduction category presents sentences expressing consequences obtained by reasoning. The most part of deductive sentences in biological literature establishes a link between the *results* and the *conclusions/demonstrations* categories.

Examples:

(9) "Furthermore, genetic and transplantation studies indicate that both Neur and Mib act in a non-autonomous manner [18,21,22,23,25,29], indicating that endocytosis of Dl is associated with increased Dl signalling activity."

(10) "*It can be deduced that the erythroid ALAS precursor protein has a molecular weight of 64.6 kd, and is similar in size to the previously isolated human housekeeping ALAS precursor of molecular weight 70.6 kd*."

Linguistic markers used to annotate these sentences are for example *deduce, implicate, indicate, deduction*.

4.1.6 Speculation

Speculations are proposals concerning a biological problem expressed explicitly as not certain in the paper (see the following section). In their scientific methodology, biologists can be interested in speculations about a biological entity or a subject [15]. Indeed, speculations can go beyond results and therefore highlight some incompletely demonstrated results, or allow the researchers to anticipate future discoveries. Besides, these speculations can give other ways to deal with a problem and give new experimental ideas ([16]; [17]).

Examples:

(11) "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."

(12) "This agrees with a recent report that suggested protein-protein interactions are more conserved within species than across species (49)".

Such markers can be verbs (*suppose, hypothesize, propose, assume*), adjectives (*convincing, probable, possible, conceivable*), modal verbs (*may, might, could*) and also conjunctions *(if, whether).*

4.1.7 Absence of knowledge

The absence of knowledge concerns not resolved questions or problems without proposing possible resolutions. This category is important because it gives new ideas for future research or experiments.

Examples:

(13) "*How endocytosis of DI leads to the activation of N remains to be elucidated.*

(14) "*The exact role of the ubiquitination pathway in regulating apoptosis* is still unclear."

Linguistic markers are for instance n*ot clear, unknown, unclear, remain poor, remain unknown*.

4.2 Sub-categorization

Some of these categories (*result, conclusion/demonstration, deduction and speculation*) can also be divided in sub-categories according to new or prior classes. This sub-categorization enables to distinguish between the contributions of the paper (*new*) and other author works (*prior*). We have found specific markers to perform this task. It can be for example specific tenses or voices applied to the previously mentioned verbs such as the use of the passive present perfect tense in *prior* sub-categories or the presence of some specific constructions in the sentence (for example, "*we hypothesized", "in our study"* for *new* sub-categorization). The presence of bibliographic citations enables also to categorize sentences into *prior* subcategory.

The *new* sub-categories underline the real new contribution of a paper. The *prior* sub-categories highlight what is taken into consideration by the author among the prior studies and how. Indeed, as the scientific knowledge is in perpetual evolution, the same information can be presented for example as an observation, a conclusion, or a speculation according to the author, the field of research and some new discoveries.

The categories classified according to certainty/knowledge level (see Figure 1.) are very close to the proposed model in [14]. However [14] link them only to the "*knowledge type*" and separate them from the certainty dimension which they treat apart. In our model, we do not make this differentiation and we keep therefore a correspondence between the degree of certainty and the knowledge type because according to us these notions are intrinsically linked in the biologists approach. Therefore certainty/knowledge level categories give indications about authors trust in their statements and/or about the means used by researchers to achieve them. In addition, it seems to us more interesting to extract agreements or disagreements between *results, conclusions/demonstrations, deductions* or *speculations*.

For example, sentences (15) and (16) express respectively a disagreement and an agreement between a prior result and a new result.

(15) "*In contrast to previous results obtained using polyclonal antiseras to detect Pan/E2A proteins, we report comparable levels of Pan proteins in GH/PRL- and insulin-producing, B- and T-lymphocyte cells."*

(16) "This observation is consistent with the detection of normal CD40-induced monocyte activation in patients with CD40 ligand+ hyper IgM syndrome in whom a defect in CD40-induced B cell activation has been reported."

5. TOWARDS AUTOMATIC PRODUCTION OF THEMATIC SHEETS

A study of some markers context is important in the annotation process since it allows to remove some marker ambiguities. Indeed, the marker analysis (see Figure 1 and Table 1) shows that some of them are present in more than one category.

In the sentences (17) and (18), the same marker (the pattern "*remains unknown"*) express two different semantic values: In (17), it is a speculation but in (18) it is an absence of knowledge. This ambiguity can be difficult to resolve just by patterns and the context has to be checked in order to find eventually other markers. These markers can be in some cases very distant in the sentence from the first marker "*remains unknown*" Thus, the presence of "*whether*" indicates that the sentence (17) is a speculation whereas the presence of "*how*" indicates that the sentence (18) expresses an absence of knowledge.

(17) "*As a consequence, it remains unknown whether the phenomenon of reduced elimination by sand filters over time due to accumulation could apply to these pathogens as well."*

(18) "*How endocytosis of DI leads to the activation of N remains unknown*."

The Contextual Exploration processing [18] is the accurate method to solve these ambiguities and to well annotate the presented semantic categories. It is a linguistic and computational method implemented in the EXCOM-2 platform ([19], [20]) that allows the annotation of segments (which can be a title, a paragraph, a sentence or a clause) according to a given viewpoint (*definition, citation, results, hypothesis…).*

EXCOM-2 uses declarative rules built by linguists or domain experts and are triggered off by the presence of linguistic markers in a text. These markers are either indicators or clues (both expressed into regular expressions).

The clues are used to confirm, invalidate or specify an annotation carried by an indicator since, sometimes, the presence of an indicator requires looking for supplementary markers.

EXCOM-2 segments automatically texts into paragraphs and sentences and then starts the automatic annotation process.

This processing is presented in Figure 2 where IND is an indicator that belongs to a semantic category and CL1, CL2…CLn are the relative clues of the called rule. The successive steps for the automatic text annotation are:

- Search for indicators of one or few given semantic categories in the segment.

- Call and execution of the associated contextual rule which are triggered by the identification of an indicator IND in the sentence (the syntax rule is shown in Figure 2).

- Search for clues (CL1, CL2,.., CLn) 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.

- Semantic annotation of the segment if all the rules conditions are satisfied.

This method has the advantage to be computationally fast due to the absence of morpho-syntactic analysis and the hierarchy between indicators and clues (this avoids the system to search for the same pattern *"remains unknown"* few times in order to categorize for example the sentences (17) and (18)).

The Contextual Exploration processing already enabled to develop the TNT-EXCOM application [21]. It allows not only to extract the non textual elements of a paper but also to establish a link with their comments in the text. As the Contextual Exploration has been proved to be able to perform this kind of task, we have now to adapt this methodology to the specificity of biological texts.

In order to illustrate the efficiency of the Contextual Exploration for annotating biological texts, we describe here the current progress of the BioExcom project that is already able to detect speculation. For this task, BioExcom uses twelve indicator classes (same semantic or grammatical categories) and thirty rules [22].

For example, the use of Contextual Exploration processing enables to remove some indicator ambiguities such as the indicator *"could"* that expresses either the past form of the auxiliary verb "*can*" or its conditional form. A sentence containing the marker "*could*" is only annotated as speculation when some positive clues expressing conditionality or possibility (such as *"if", "whether*" or *"alternatively"*) are present in the indicator context. This method is able to correctly identify some speculative sentences using the "*could*" marker but is not able to take off all the ambiguities [22].

The sentence (19) is an example of a speculative sentence marked by the *"could"* indicator.

(19) "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."

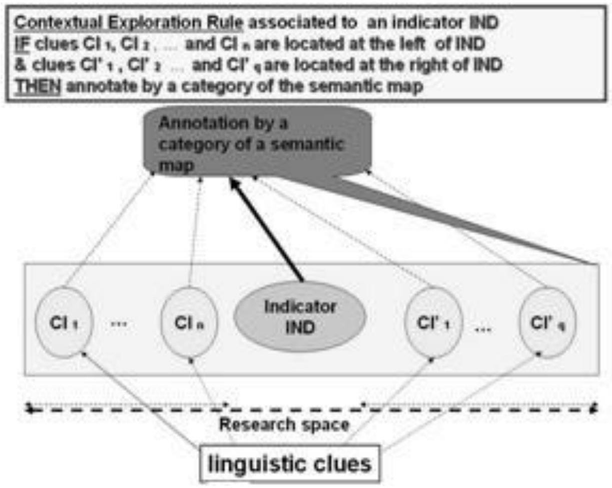
Contextual Exploration Rule associated to an indicator IND
IF clues Cl 1, Cl 2, ... and Cl n are located at the left of IND
& clues Cl' 1 , Cl' 2 ... and Cl' q are located at the right of IND
THEN annotate by a category of the semantic map
Annotation by a category of a semantic map
Cl 1
...
Cl n
Indicator IND
Cl' 1
...
Cl' q
Research space
linguistic clues

Figure 2. The contextual Exploration principles: search for an indicator (IND) and then for some clues (Cl) in a research space (the same sentence in our case) according to some associated rules.

An evaluation of the automatic annotation of speculative sentences has been already performed and the results are good [23]. Especially, an evaluation on 14 500 sentences belonging to a part of the Bioscope corpus [24] and which were manually re-annotated according to the semantic criteria of BioExcom by following a methodology described in [23] and based on the comparison between the initial BioScope annotation and the automatic annotation of BioExcom[4], was accomplished. BioExcom annotated this corpus consisting of 2688 full papers sentences (341 sentences annotated as speculation in 294 seconds) and 11 812 abstract sentences (1489 sentences annotated as speculation in 153 seconds). According to this study, BioExcom presents a F-score of 90,1 % (82,7 % recall and 99,1 % precision) for the detection of speculations, confirming the method effectiveness. An other evaluation of the categorization of speculation by BioExcom into prior and new speculation has been also realized. It has been performed on three papers including 71 speculative sentences in total by using evaluator judgment concerning the BioExcom results [22].The obtained F-score is 88,6% (84,6 % recall and 89,7% precision).

These results are not completely comparable with the previous results since they were not performed on the same scale and as the methodologies were not the same. However, they point out that the categorization process is efficient and can be applied for all other categories.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we focused on the usefulness for a biologist of finding in literature different kinds of general information, independently of biological sub-domains. More particularly, we highlighted the importance of separating between the type of knowledge (speculation, deduction and result for example) and their temporal characteristics (new or prior). We proposed a model of semantic and discursive categories with the aim to automatically create thematic sheets of biological papers by the BioExcom tool. These thematic sheets can be used for a system of automatic summarization and information extraction.

BioExcom is already able to automatically identify speculative sentences and to categorize them into new or prior speculation by using the Contextual Exploration processing. Our future work will focus on the construction of Contextual Exploration rules by using the linguistic resources in this study here and corresponding to the categories described here.

This work will enable the automatic annotation of biological texts and then the development of a user interface connected to biological named entity dictionaries in order to allow to the researcher to find specific information. The ultimate goal is to propose to researchers a personal and semantic bibliographic management tool for full papers.

[4] The corpus BioSpe manually annotated for speculation is available at http://www.bioexcom.net/.

7. REFERENCES

[1] Ananiadou, S., Kell, D.B., Tsujii, J. 2006.Text mining and its potential applications in systems biology. *J. Tre. Bio.* 24 (Oct 2006), 571-579.

[2] Jensen, L.J., Saric, J., Bork, P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet 7,* 119-129 (Feb 2006).

[3] Rzhetsky, A., Seringhaus, M., Gersein, M. 2008. Seeking a new biology through text mining. *J.Cel.* 134 *(Jul 2008)*, 9-13

[4] Cohen, B., Hunter, L. 2008 Getting started in text mining , Plos Computational biology volume 4, issue 1 (Jan 2008)

[5] Hu, Z.Z., Narayanaswamy, M., Ravikumar, K.E., Vijay-Shanker, K., Wu, CH. 2005. Literature mining and database annotation of protein phosphorylation using a rule-based system. *J.Bio.* 21 (Apr. 2005), 2759-2765

[6] Jensen, L.J., Saric, J., Ouzonova R., Rojas, I., Bork, P. 2006. Extraction of regulatory gene/proteína Networks from Medline. *Bioinformatics*,Vol. 22 no. 6 2006, pages 645–650 *doi:10.1093/bioinformatics/bti597*

[7] Wilbur W.J., Rzhetsky A., Shatkay H. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *J.BMC.Bio*.7 (Jul 2006), 356

[8] Teufel, S., Moens, M. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *J.Comp Lin, 28*,4 (Dec 2002).

[9] Mizuka, Y., Korhonen, A., Mullen, T., Collier, N. 2006. Zone analysis in biology articles as a basis for information extraction. *J. Med. Inf.* 75 (Jun 2006), 468-487.

[10] Blais A., Atanassova I., Desclés J.P., Zhang M., Zighem, L. 2007. Discourse Automatic Annotation of Texts: an Application to Summarization. In *Proceedings of the Twentieth International FLAIRS Conference, Special Track «Automatic Annotation and Information Retrieval: New Perspectives»* (Key West Florida, USA, May 7-9, 2007). FLAIRS„07

[11] Harmsze, F.A.P. 2002 *Modular structure for scientific articles in an electronic environment*. Doctoral Thesis, Univérsity of Amsterdam.

[12] Waard, A. de., Buitelaar, P., Eigner T. 2009 Identifying the Epistemic Value of Discourse Segments in Biology Texts, In *Proceedings of the 8th International Conference on Computational Semantics (*Tilburg, Netherland; January 7-9, 2009*)*, 351-354. DOI= http://portal.acm.org/citation.cfm ?id=1693756.1693802&coll=DL&dl=GUIDE&CFID=115124777 &CFTOKEN=99232089

[13] Rubin, V. L., Liddy, E. D., Kando, N. 2005.Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In J. G. Shanahan, Y. Qu & J. Wiebe (Eds.), Computing Attitude and Affect in Text: Theory and Applications (the Information Retrieval Series). New York: Springer-Verlag, 61-76.

[14] Thompson, P., Venturi, G., McNaught, J., Montemagni, S., Ananiadou, S. 2008. Categorising modality in biomedical texts. In *Proceedings of the Workshop on Building and Evaluating 1 Resources for Biomedical Text Mining*. (Marrakech, Morocco, May 26-30, 2008) *LREC'08.*

[15] Light M., Qiu X.Y., Srinivasan P. 004.The Language of Bioscience: Facts, Speculations, and Statements in Between. In *Proceedings of HLT-NAACL Workshop On Linking Biological Literature Ontologies And Databases*, 17-24 (Boston, Massachusetts May 6, 2004). HLT-NAACL‟ 04.

[16] Bray, D. 2001. Reasoning for results. J.*Nat.* 412 (Aug 2001), 863

[17] Blagosklonny, MV., Pardee, AB. 2002. Conceptual biology: unearthing the gems. *J.Nat.* 416 *(Mar 2002),* 373.

[18] Desclés, J.P. 2006. Contextual Exploration Processing for Discourse Automatic Annotations of Texts. In *Proceedings of the Nineteenth Florida Artificial Intelligence Research Society International conference*, (Melbourne Florida, USA, May 11_13, 2006) 281-284. FLAIRS‟ 06.

[19] Djioua, B., Flores, J.G., Blais, A., Desclés, J.P., Guibert, G., Jackiewicz, A., Le Priol, F., Nait-Baha, L., Sauzay, B. 2006. EXCOM: an automatic annotation engine for semantic information, In *Proceedings of the Nineteenth Florida Artificial Intelligence Research Society International conference*, (Melbourne Florida, USA, May 11_13 2006) 285-290. FLAIRS‟ 06.

[20] Alrahabi, M., Desclés J.P. 2009. EXCOM: Plate-forme d'annotation sémantique de textes multilingues, In Proceedings of the Natural Language Processing conference (Senlis, Canada, June 24-26 2009). TALN‟ 09.

[21] Le Priol, F. 2008. Automatic annotation of images, pictures or videos comments for text mining guided by no textual data, In *Proceedings of the twenty first Florida Artificial Intelligence Research Society International conference* (Miami Florida, USA, May 15-17) FLAIRS‟ 08.

[22] Desclés, J., Alrahabi, M., Desclés J.P., Blais A. 2009 Automatic Annotation by BioExcom for categorizing prior and new speculations in biological papers. Presented at *3rd International Symposium on Languages in Biology and Medicine*, (Jeju Island, South Korea, November 8-10 2009). LBM‟ 09

[24] Desclés, J., Makkaoui, O., Hacène, T. 2010. Automatic annotation of speculation in biomedical texts: new perspectives and large scale evaluation. In *Proceedings of the Negation and speculation in Natural Language Processing workshop,* (Uppsala, Sweden, July 10, 2010). NeSpNLP‟ 10 DOI=http://portal.acm.org/citation.cfm?id=1858959.1858965&coll=DL&dl=GUIDE&CFID=115124777&CFTOKEN=99232089.

[25] Szarvas, G., Vincze, V., Farkas, R., Csirik, J. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing.* BioNLP ACL‟ 08, (Colombus, Ohio, Jun 19-20, 2008) DOI= http://portal.acm.org/citation.cfm?id=1572306.1572314&coll=DL &dl=GUIDE&CFID=115349241&CFTOKEN=40214864