CHAPTER 2

Constructing Synthesized Sheets by Mining Scientific Research Papers: Application to the Biological Domain

Olfa Makkaoui, Leila Makkaoui, Iheb Kechaou and Jean-Pierre Desclés

This chapter presents a text mining tool for scientific publications that allows the extraction of textual segments (section, paragraph, sentences, etc.) from a large corpora according to a set of semantic categories (results, methods, hypothesis, etc.). The extracted information is grouped according to their semantic affiliation which allows to obtain an organized textual representation called multi-document synthesized sheets. The automatic construction of these synthesized sheets is realized by semantically annotating documents according to a set of semantic categories. In fact, the annotation task is performed automatically using the Contextual Exploration processing (EC). It

Olfa Makkaoui and Jean-Pierre Desclés
LaLIC - Université Paris-Sorbonne28
rue Serpente 75006 Paris, France
e-mail: olfa_makkaoui@yahoo.fr, jean-pierre.descles@paris-sorbonne.fr

Leila Makkaoui
Institut du Cerveau et de la Moelle épinière (ICM)
e-mail: leilamakkaoui@gmail.com

Iheb Kechaou
Radiologie Diagnostique & Interventionnelle, Hopital Marie Lannelongue
e-mail: iheb.kechaou@yahoo.fr

Editors: S. Murala and S.K. Vipparthi, *Multimedia Retrieval - Perspectives and Challenges*
DOI: 10.15579/gcsr.vol4.ch2, GCSR Vol. 4, pp. 19-42, 2015

is a computational linguistic method based on a set of linguistic markers associated with semantic categories.

2.1 Introduction

Scientific publications contain results and ideas that can point to possible new discoveries. We propose in this context an automatic text mining tool that annotates automatically scientific publications according to semantic categories (results, method, hypothesis, etc.) and classifies them into synthesized sheets. These are considered as an organized and structured representation of textual segment (sentences).

Mutli-document synthesized sheets allow:

- Crossing information from different documents which can help to discover new knowledge by extracting information from different research papers and/or area.
- Access to the information extracted and categorized according to the user's choice. He can extract the most relevant information found in a corpus regarding a studied subject.
- Building structured thematic summaries of scientific papers depending on specific kinds of information (result, hypothesis, method, etc.). This enables users to get rapidly the real output of scientific papers.

The automatic construction of synthesized sheets requires the automatic and semantic annotation of the textual document. This task is performed by the Contextual Exploration processing. It is a computational linguistic method based on a set of linguistic markers associated with semantic categories.

We aim in this chapter to present:

- The automatic process that enables the annotation of scientific publications. This methodology is performed in biological literature.
- A user interface that enables the automatic construction of synthesized sheets.

2.2 Synthesized Sheets to Mine Scientific Papers

The construction of multi-document synthesized sheets is based on the semantic annotation of documents according to a set of semantic categories called *"semantic map"* that may interest users. In order to determine them, we conducted a linguistic analysis of thirty biological papers ([9]. This study allows the extraction of a set of linguistic markers for each semantic category. These linguistic marker sets are enriched using synonym dictionaries. We propose two types of categories (Fig.(2.1)):

- General information: It contains semantic categories that deal with the general ideas of the paper ("Thematic announcement", "Method" or "Text/image association").

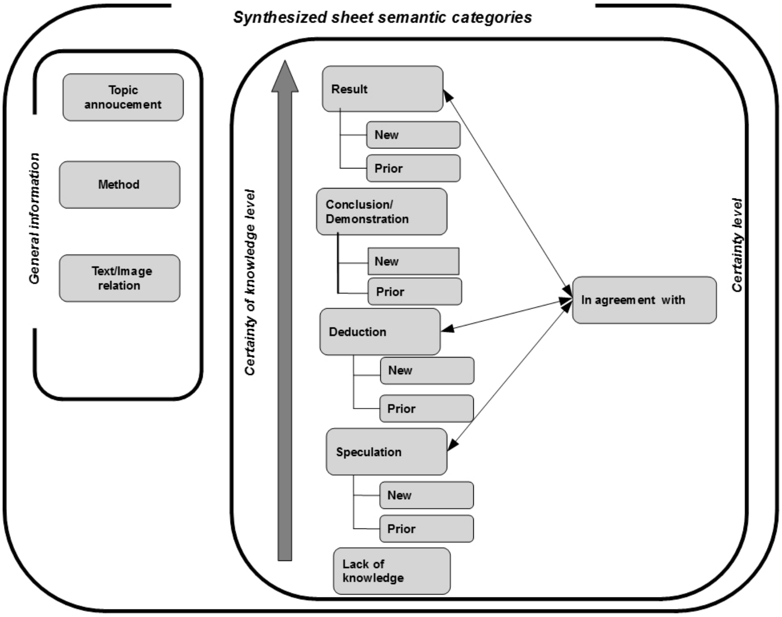


Figure 2.1: Semantic categories of synthesized sheets.

- Certainty level based categories: These categories depend on the author's reliability regarding the communicated information. They can be ordered from unreliable to certain: "Absence of knowledge", "Speculation", "Deduction", "Conclusion", "Result".

2.2.1 General Information

Thematic announcement
This semantic category gives an idea about the topics covered in the text.
Examples:

(1) *"**We present the first systematic study** of their structure via synchrotron X-ray computed tomography and high-resolution scanning electron microscopy".*

(2) *"**The aim of this paper** is to present the BioTRON system, which supports biologists in the various steps necessary to perform complex biological tasks such as biological network comparison".*

Method
This category details methodologies used in the research described in the paper.
Examples:

(3) *"Sypro-Ruby dye (Bio-Rad) **was used** to quantify the amount of proteins in the complex".*

(4) *"Ligase IV knockdown* ***was performed*** *with siRNA or antisense Ligase IV plasmid by transfecting into MCF7, HeLa, and Nalm6 cells with oligofectamine and lipofectamine (Invitrogen), respectively, whereas over expression was performed as per standard protocol".*

Text/image association

Scientific publications contain many non-textual elements (figures, images, etc.). Focusing only on extracting textual segments do not always give complete information while non-textual segments contain detailed and complete data.

We propose to construct non-textual synthesized sheets by extracting non textual elements like figures and linking them to their comments in the text. This task is already presented in [30]. It should be also noted that many results or methodologies are presented in figures. For example, in the sentence (5) authors describe the obtained results by means of a figure.

(5) *"The 30Si results are presented in Fig.2"* .

In this case, we propose in addition to the annotation of this sentence as a "result", to display the figure mentioned in the synthesized sheet which enables users to get the complete information.

Examples:

(6) *"In order to facilitate and harmonize the approaches to understand the role of Si in planta, Ghanmi et al.[46] proposed the use of the Arabidopsis-powdery mildew interaction, by showing that this model plant reacted to powdery mildew as other dicots and monocots did under Si treatment (Fig. 2)".*

(7) *"The 30Si results are presented* ***(in Fig. 2 and Table 1)****".*

Figure 2.2 shows a synthesized sheet of non-textual elements extracted by our system (more details about the user interface is presented in Section 2.5).

2.2.2 Certainty Level Based Categories

Result

This category deals with results obtained or described in the paper.

Examples:

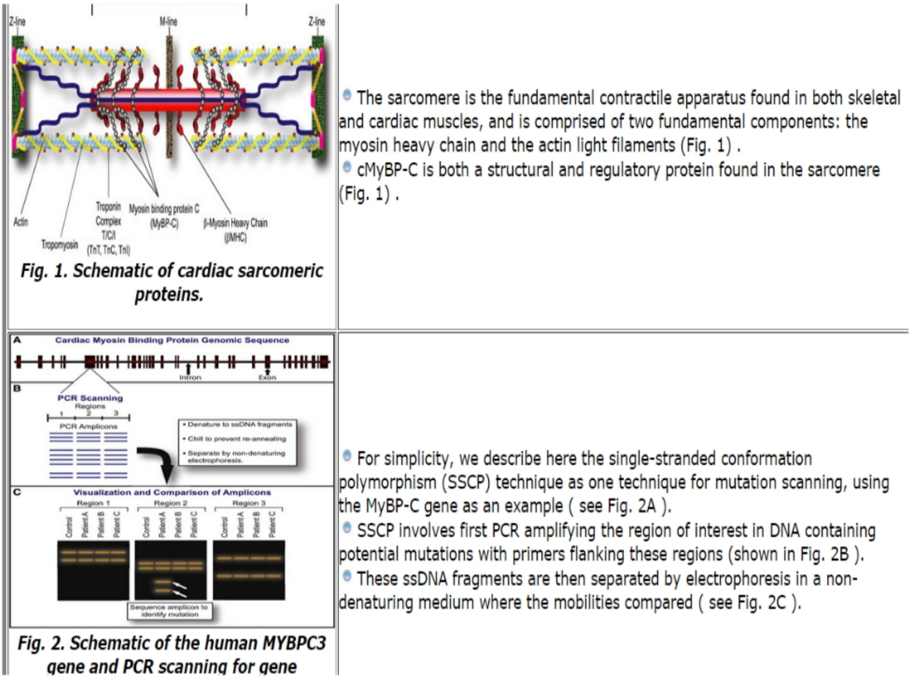
(8) *"A recent study of RXR complex to 9-cis retinoic acid and SRC-1 NR2* ***revealed*** *a difference in the conformation of the 9-cis RA in the coactivator bound RXR (32)"* .

(9) *"****Our results showed that*** *when treated along with SCR7, ionizing radiation (IR) and etoposide could enhance tumor regression more efficiently".*

Conclusion

Conclusions are demonstrations presented in the papers. This semantic category can also summarize information that deal with the whole paper.

Examples:



Z-line
M-line
Z-line
Actin
Tropomyosin
Troponin Complex T/C/I (TnT, TnC, TnI)
Myosin binding protein C (MyBP-C)
β-Myosin Heavy Chain (βMHC)
A
Cardiac Myosin Binding Protein Genomic Sequence
Intron
Exon
B
PCR Scanning
Regions
1 2 3
PCR Amplicons
• Denature to ssDNA fragments
• Chill to prevent re-annealing
• Separate by non-denaturing electrophoresis.
C
Visualization and Comparison of Amplicons
Region 1
Region 2
Region 3
Control
Patient A
Patient B
Patient C
Sequence amplicon to identify mutation

Fig. 1. Schematic of cardiac sarcomeric proteins.

- The sarcomere is the fundamental contractile apparatus found in both skeletal and cardiac muscles, and is comprised of two fundamental components: the myosin heavy chain and the actin light filaments (Fig. 1) .
- cMyBP-C is both a structural and regulatory protein found in the sarcomere (Fig. 1) .

Fig. 2. Schematic of the human MYBPC3 gene and PCR scanning for gene

- For simplicity, we describe here the single-stranded conformation polymorphism (SSCP) technique as one technique for mutation scanning, using the MyBP-C gene as an example (see Fig. 2A).
- SSCP involves first PCR amplifying the region of interest in DNA containing potential mutations with primers flanking these regions (shown in Fig. 2B).
- These ssDNA fragments are then separated by electrophoresis in a non-denaturing medium where the mobilities compared (see Fig. 2C).

Figure 2.2: Image quality comparison of original image (first column) and EDBTC reconstruction (second column) in RGB color space. The third and fourth column are the original and EDBTC image reconstruction, respectively, in YCbCr color space.

(10) *“We concluded that the b-isox chemical itself forms microcrystals that selectively coprecipitate RNA-binding proteins containing LC domains”.*

(11) *“In the present study,* ***we demonstrate*** *that binding of the coactivator to one RAR subunit exerts an allosteric control over its own interaction with the second RAR protomer”.*

Deduction

Deductions are consequences obtained from reasoning resulted from used methodologies.

Examples:

(12) *“The structural superposition of the crystal structure of monomeric RARb − TTNPBLBD (PDB ID code 1XAP) onto M1 and M2 subunits of the homodimer* ***indicates that*** *monomer M1 is much closer to the monomeric conformer (rmsd 0.46 versus 0.65 for M2)”.*

(13) *“We* ***deduce that*** *the biologically active conformation at the gastrin receptor is partly helical and one in which the indole of tryptophan and the aromatic ring of phenylalanine are close to one another while the methionine and aspartic acid side chains point in the opposite direction”.*

Speculation

Recent research in text mining linked to the biological domain has made major progress and took into consideration the importance of extracting speculation by distinguishing between factual statements and uncertainty [18, 24]. This task is especially linked to the consideration that biological researchers can be only interested in finding factual sentences in the text. Information is consequently classified as facts or speculations. These latter are considered in this case as hedges since their meaning is general and concerns all information that do not belong to the factual statements. However, biologists can be also interested in extracting speculations linked for example to a particular entity. This task is important for their experimental research as authors are not sure about their results and the speculations they provide can be a starting point for new experiments [21, 6].The meaning of speculation is in this case more restrictive than hedges and is very close to hypothetical statements. This latter speculation characterization is developed by our system that aims to answer to the biologists needs concerning the extraction of speculative sentences in biological texts [9]. This approach underlines the importance of establishing a link between their experimental findings and ideas or proposals about biological issues provided in the literature without taking into account approvals or negations of them. Our aim is to extract from the scientific literature, ideas and proposals about a particular topic by considering speculation as a potential source of information. In [9] a speculation is defined as a non-demonstrated proposal about a biological problem which is explicitly presented as uncertain in the paper. This information is important as it can highlight knowledge not yet demonstrated and anticipate future experiments.

Examples:

(14) *“The SOFeXN export event was* ***probably*** *triggered by the subduction of phytoplankton to depth when the bloom filament encountered a front (10)”.*

(15) *"Proteins* ***may*** *be unfolded, partially unfolded or native (Chilson & Chilson, 2003)"* .

Absence of knowledge

The absence of knowledge category deals with problems and questions not resolved in the described research.

Examples:

(16) *"How specific RNA-binding proteins are chosen for inclusion in RNA granules remains unclear".*

(17) *"However, little is known about inhibitors against core NHEJ proteins, such as KU70/80 complex, Artemis, Ligase IV/XRCC4, Pol μ, and Pol λ".*

2.2.3 Sub-categorization and Relation Between Semantic Categories

Most of the certainty level based categories (all categories except "Absence of knowledge") are sub-categorized into *«new»* and *«prior»* sub-categories:

- The prior subcategory deals with information reported by the authors.
  Example:

(18) *"Our* ***analysis show*** *that mitochondrial barrel channels from Archaeplastida".*

- The new subcategory presents the real output of the paper.
  Example:

(19) *"A recent study* ***showed that*** *by manipulating one of the genes responsible for a protein in this clock process, a behavioral change very much like maniac could be produced in a mouse".*

Certainty level categories are interlinked by a relation of agreement (in agreement / disagreement with). This relation is useful as it determines ideas that converge to similar approaches or uncover opposite statements. For example, the sentence (20) is an agreement between a reported result and a new result and the sentence (21) deals with a disagreement between a reported result and a new result.

Examples:

(20) *"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".*

(21) *"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".*

Table 2.1: Examples of linguistic markers of the semantic categories.

Semantic categories	**Example of linguistic markers**
Topic announcement	aim of this study, idea, intention, purpose, address, idea, objectif of our work
Method	use, measure, monitor, method, test, technique, perform, test
Text/image Relation	figure, fig, Picture, Illustration, image
Result	identify, result in, reveal, show, discover, find, show that
Conclusion	prove, demonstrate, state, in summary, assert, report, demonstration
Deduction	infer, indicate that, deduce, deduction, signal, consequence
Speculation	suggest, may, perhaps, could, hypothesis, possible, probable
Absence of knowledge	be not known, remains unknown, is not clear how, be an open question, still unclear

2.3 Automatic Annotation Task

In order to automatically annotate textual documents, we use the Contextual Exploration processing [12, 13]. The identification of textual segments is possible due to the presence of some linguistic markers identified in the text. Linguistic markers that enable the identification of semantic categories are called «indicators». These are domain-independent and can be applied in other domains such as sciences or sociology. We present in Table 2.1 some examples of linguistic markers of each category of the semantic map:

The presence of an indicator in a textual segment indicates the possibility of its annotation for a particular semantic category. However, most of the indicators are ambiguous and their presence in a textual segment does not automatically imply its affiliation to a semantic category. For example, although both of the two following sentences (sentence (22) and (23)) use the same indicator, they express two different meanings. Indeed, the presence of the "whether" clue indicates that the sentence (22) is a speculation whereas the "how" clue shows that the sentence (23) expresses an absence of knowledge. This latter notion deals with open questions without presenting any proposal or idea about a subject:

(22) "Also, whether the signaling activity of Ser is similarly regulated by endocytosis is not known".

(23) "How endocytosis of DI leads to the activation of N is not known".

As in some cases the simple detection of these indicators is not sufficient to correctly annotate sentences, the Contextual Exploration processing focuses on some additional linguistic markers (clues) in the indicator context in order to remove ambiguities. Linguistic clues can be positive if they enable to confirm an annotation decision or negative if they are used to contradict it.

Successive steps for the application of the Contextual Exploration processing are (Fig.(2.3)):

- Step 1: Looking for indicators of one or few given discursive categories in the segment.
- Step 2: Call and execution of the associated Contextual Exploration rules which are triggered by the identification of an indicator in the sentence.

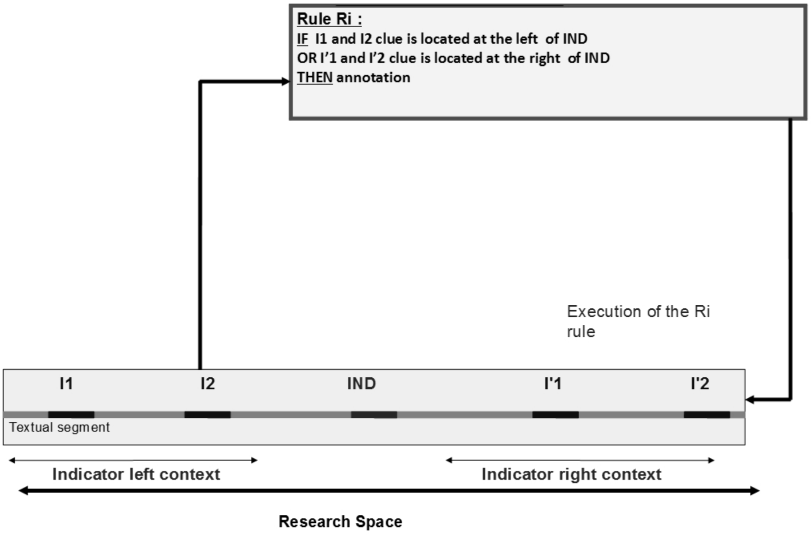


Figure 2.3: The Contextual Exploration principles: search for an indicator (IND) and then for some clues (I1, I2, I'1, I'2...) in a research space (the same sentence in our case) according to some associated rules.

- Step 3: Looking for clues contained in the rule. These clues can be performed in the sentence at the right or/and the left of the indicator or even inside the indicator.

- Step 4: Semantic annotation of the segment if all the rules condition are satisfied.

The Contextual Exploration processing is performed by the EXCOM-2 annotation platform [1, 13] that enables to automatically annotate texts according to a given semantic category in successive steps. First, texts are segmented into sentences using a list of typographical signs. The obtained segments (which can be sections, paragraphs or sentences) are then automatically annotated using the Contextual Exploration processing.

The categorization of speculative sentences into "*new*" and "*prior*" subcategories task was based on the search for some specific verbal aspects and also specific linguistic clues (additional markers):

Additional linguistic markers (clues) used in the "*prior*" subcategory:

- The presence of bibliographic citations in the sentence as positive clues:

(24) "*Chretienet al (1995)* ***hypothesize that*** *closure of protofilament sheets triggers GTP hydrolsis*".

(25) "*This method **demonstrated** importance of polymerization dynamics to MT function during mitosis [1]*".

- The presence of specific expressions that indicate that the author is presenting other works such as "*recently*", "*in their work*", "*previous studies* ".

(26) "*A and water as solvent B, these solvents used a standard linear gradient, as **demonstrated** in their paper, at 32o C*".

(27) "*Their experiments **revealed** the presence of heat-resistant bacteria whose inactivation required raising the can center to a minimum of 250 degrees Farenheit and holding it there for at least 10 minutes*".

Additional linguistic markers (clues) used in the "*new*" subcategory:

- The presence of specific expressions considered as positive clues such as "*here*", "*our analysis*".

(28) "*Here we have **assumed that** premature termination of translation does not play a dominant role*".

(29) "*This work **demonstrates that** different types of nanostructured materials were efficiently used as hosts for enzyme*".

- The absence of bibliographic references in the sentence. According to us, it indicates that the author is not presenting other works. In this case, bibliographic references are considered as negative clues and their presence in the indicator context invalidate the possibility of its annotation as a "*prior*" subcategory.

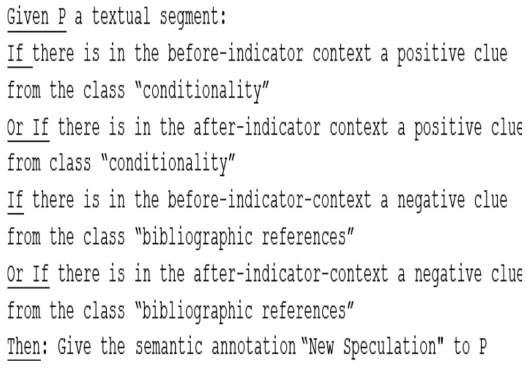
(30) "*By analogy with Ras, which acts as a signalling molecule, Ran-GTP **might be** the"active" form that binds effecttors*".

(31) "*Biochemical information **indicates that** the signaling activity of Ran is determined by its GTP- or GDP-bound state*".

Figure 2.4 illustrates an example of the application of a Contextual Exploration rule concerning the sentence (22):

(32) "*Also, whether the signaling activity of Ser is similarly regulated by endocytosis **is not known***".

The indicator "*is not known*" is first detected and consequently triggers the execution of the Contextual rule associated to this indicator. Thus, according to the executed rule, additional markers are searched in the indicator context to confirm or infirm the annotation. Positive clues are expressions of conditionality like "*if*" and "*whether*" and their presence confirms that the sentence can be annotated as "*speculation*". Negative clues are Bibliographic references (for example *[2], [Fritsh, 1989], (Fritsh et al. 1989)*, etc.) and their absence confirms that the sentence is a "*new speculation*".



```
Given P a textual segment:
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"
If there is in the before-indicator-context a negative clue
from the class "bibliographic references"
Or If there is in the after-indicator-context a negative clue
from the class "bibliographic references"
Then: Give the semantic annotation "New Speculation" to P
```

Figure 2.4: Example of the Contextual Exploration rule (the used indicator is "*not known*" and the annotation action is: "*new speculation*").

Table 2.2: Summary of raw results for the evaluation.

	Precision	**Recall**	**F-Measure**
Full Text	89.35%	62.92%	73.84%
Abstracts	94.75%	68.83%	79.74%

2.4 Evaluation

We present in this section the evaluation process of the annotation of the semantic categories proposed to mine scientific papers. The speculation detection task was first evaluated on a small corpus and enabled to prove the method's effectiveness [9]. The evaluation of the speculation category is already realized on a large corpus [11, 23]: The BioScope corpus[33]. It consists of three parts namely medical free texts, biological full papers and biological scientific abstracts. Only the biological full papers and the biological scientific abstracts parts (consisting of 9 fulltexts and 1273 abstracts) of the BioScope corpus were analyzed because we are especially interested in the biomedical scientific domain. The annotations tags of the two BioScope corpus parts were first removed then automatically segmented and annotated by our system. The latter automatically annotated 1830 sentences (341 sentences from full text papers and 1489 sentences from the abstracts corpus part). The categorization into "*new*" and "*prior*" speculation was not taken into consideration during the evaluation process.

The evaluation results are presented in Table 2.2. The Precision is approximately 93% in average (calculated from the total of segments of the two corpora) and the Recall is approximately 68% (in average).

Our aim is to evaluate the annotation performance of the other semantic categories. This evaluation is realized on a corpus of papers selected from different journals[1]. Ten

[1] Example of journal papers: Nature, Science, Plos biology, PNAS, Cell, etc.

Table 2.3: Annotation statistics of the evaluated corpus.

	Number of Sentences	**Number of annotations**
Paper 1	357	79
Paper 2	380	140
Paper 3	398	44
Paper 4	439	53
Paper 5	348	120
Paper 6	456	150
Paper 7	291	117
Paper 8	219	148
Paper 9	269	129
Paper 10	401	131

Table 2.4: Results of the automatic annotation for each category.

Semantic categories	**Number of annotated sentences**
Thematic announcement	4
Method	208
Text/image relation	203
Result	279
Conclusion	107
Deduction	20
Speculation	255
Absence of knowledge	19

papers are randomly selected from this corpus and then segmented and annotated automatically by our system. Three evaluators read the corpus and specify if they are in agreement with our system annotation decisions (the annotation or the absence of the annotation), and have also to propose their own annotation according to the semantic categories of our system. The annotation statistics are presented in Table 2.3 and Table 2.4.

The inter-annotator is presented in Table 2.5. The evaluation results are generally good (Table 2.6). However, the analysis of the evaluated sentences reveals that our system did not detect some sentences due to the lack of some specific linguistic markers and a weak decrease of performance is also observed. For example, the sentence (32) expresses an absence of knowledge but was not detected by our system.

(33) "*The mechanism whereby mutations in a single gene that is widely expressed cause such diverse diseases remains a puzzle*".

The following sentence (sentence 33) is annotated as a prior result while it deals with new result.

(34) "*We then applied these methods to studies of human urine and plasma and*

Table 2.5: Statistics of inter-annotator agreement.

	Inter-annotator agreement
Thematic announcement	82.48%
Method	78.10%

Table 2.6: Results of the evaluation.

Semantic category	**Precision**	**Recall**
Thematic announcement	75%	50%
Method	84.13%	68.36%
Text/image relation	91.16%	78.32%
Result	91.24%	87.32%
Result categorization	88.60%	77.53%
Conclusion	93.45%	85.98%
Conclusion categorization	84.11%	76.63%
Deduction	84.95%	69.23%
Deduction categorization	66.66%	52.15%
Absence of knowledge	89.47%	78.94%

showed that *the assay was linear from 0.025 to 80 Î 14 g/L and in human plasma from 0.0025 to 80Î 14g/L (r [2] > .99)".*

2.5 Process of Construction of Synthesized Sheets

The user-interface that we introduce aims to automatically generate synthesized sheets obtained based on the intersection between the semantic annotations and the users' requests. The process of constructing the synthesized sheets (presented in the Fig.(2.5)) is based on a group of modules realized in successive steps. First, the documents that will constitute the synthesized sheets are segmented and then automatically annotated by using the EXCOM-2 annotation platform following the semantic map presented in (Fig.(2.1)). The annotated documents provided in XML are stored in a database that contains the textual segments of these documents as well as their corresponding annotations.

The proposed interface makes a distinction between two types of synthesized sheets:

Single document synthesized sheets:

These synthesized sheets enable the presentation of the annotated sentences that resulted from a single document. They are regrouped according to the discursive category of the proposed semantic map. The navigation between these categories provides the reader with a general idea of the main ideas of the paper as well as an

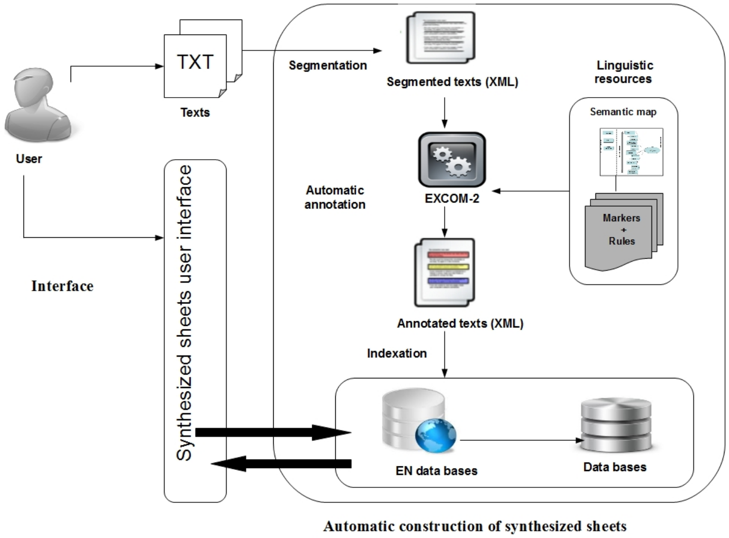


Figure 2.5: User Interface for the automatic construction of synthesized sheets.

access focused on the results of the research described. The interface dedicated to the construction of synthesized sheets, made up of a single document, is presented in the figure below (Fig.(2.6)).

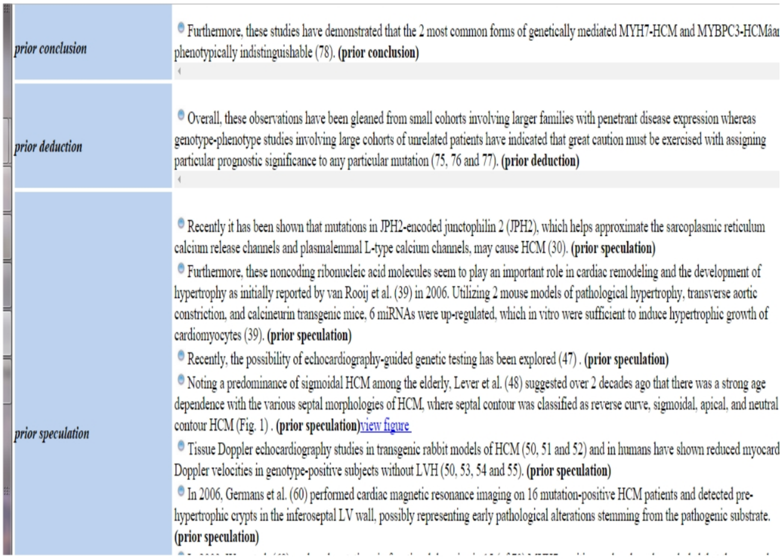
Multiple documents synthesized sheets:

The multiple documents synthesized sheets assemble annotation based on multiple texts or corpora. In this case, the process of construction of the synthesized sheets can be guided by the selection of certain semantic categories or/and the selection of one or more terms.

Indeed, the selection of semantic categories is useful as a user can focus on certain semantic categories more than on others. For example, he can be interested in the identification of annotated sentences (for example "*Topic announcement*") within a corpus of documents. This provides a general idea about the treated subjects and predicts which article potentially deserves to be read.

We aim, by the selection of one or more terms, to refine the results of the construction of synthesized sheets by means of selecting one or more terms in addition to the choice of semantic categories. This functionality is useful for searching and retrieving information regarding a particular subject. For instance, a biology researcher can be interested in the generation of a synthesized sheet linked to the results *new result* and *prior result* about a biologic entity like *NF-KB*. Nevertheless, this gene can be defined by multiple terms such as:

- DNA-binding factor KBF1
- NF-kappaB



prior conclusion	Furthermore, these studies have demonstrated that the 2 most common forms of genetically mediated MYH7-HCM and MYBPC3-HCMâaı phenotypically indistinguishable (78). **(prior conclusion)**
prior deduction	Overall, these observations have been gleaned from small cohorts involving larger families with penetrant disease expression whereas genotype-phenotype studies involving large cohorts of unrelated patients have indicated that great caution must be exercised with assigning particular prognostic significance to any particular mutation (75, 76 and 77). **(prior deduction)**
prior speculation	Recently it has been shown that mutations in JPH2-encoded junctophilin 2 (JPH2), which helps approximate the sarcoplasmic reticulum calcium release channels and plasmalemmal L-type calcium channels, may cause HCM (30). **(prior speculation)** Furthermore, these noncoding ribonucleic acid molecules seem to play an important role in cardiac remodeling and the development of hypertrophy as initially reported by van Rooij et al. (39) in 2006. Utilizing 2 mouse models of pathological hypertrophy, transverse aortic constriction, and calcineurin transgenic mice, 6 miRNAs were up-regulated, which in vitro were sufficient to induce hypertrophic growth of cardiomyocytes (39). **(prior speculation)** Recently, the possibility of echocardiography-guided genetic testing has been explored (47) . **(prior speculation)** Noting a predominance of sigmoidal HCM among the elderly, Lever et al. (48) suggested over 2 decades ago that there was a strong age dependence with the various septal morphologies of HCM, where septal contour was classified as reverse curve, sigmoidal, apical, and neutral contour HCM (Fig. 1) . **(prior speculation)**view figure Tissue Doppler echocardiography studies in transgenic rabbit models of HCM (50, 51 and 52) and in humans have shown reduced myocard Doppler velocities in genotype-positive subjects without LVH (50, 53, 54 and 55). **(prior speculation)** In 2006, Germans et al. (60) performed cardiac magnetic resonance imaging on 16 mutation-positive HCM patients and detected pre-hypertrophic crypts in the inferoseptal LV wall, possibly representing early pathological alterations stemming from the pathogenic substrate. **(prior speculation)**

Figure 2.6: Visualization of a single document synthesized sheet.

- EBP-1
- Nfkb1
- Nuclear factor kappa B p105 subunit

This example shows that the identification of biologic entities within the texts can prove to be quite complex. Actually, narrowing the results of a search based on an exact correspondence with the terms they contain does not allow the retrieving of all the possible results. Therefore, the multitude of terms attributed to a single name of gene can have negative consequences on the results of the research and on the extraction of information, such as the gathering of irrelevant information (precision base) or the negligence of an important amount of results (decrease of the reminder level) [32].

In order to demonstrate the impact of the variation of terms on the results of the research, we will present a study that focuses on the gene *JNK* conducted by Wren [37]. This study consists in the research of the apparitions of this gene (*JNK*) in the databases (PUBMED[2] and OVID[3]).

These are the different terms used for referring to this gene:

- JNK
- c-jun N-terminal kinase

[2] http ://www.ncbi.nlm.nih.gov/pubmed
[3] http://www.ovid.com

- c-jun NH2-terminal kinase
- c-jun amino-terminal kinase
- jun N-terminal kinase
- MAPK8 (official LocusLink name, ID#5599)
- Mitogen activated protein kinase

The analysis of these different forms of the entity *JNK* emphasizes certain properties that characterize the biological entities:

Synonymy:

This property is considered one of the most remarkable specificities of these entities. Basically, it aims at defining a single entity by using different terms. It is believed that certain genes belonging to the domain of *Flybase*[4] can reach up to 11 synonyms. But in the scientific literature, the probability that two authors use the same term for the same entity is less than 20%.

Terms Variation:

The entities belonging to the domain of biology have multiple formats. Most of these entities are not constituted by a single word but rather made up by several terms such as "*Mitogen activated protein kinase*". Nenadic et al. [28], analyse the biologic entities present in the corpus *GENIA* [19] and demonstrate that 85.07% of the biologic entities present in this corpus are constituted by several words. This percentage can reach up to 90% in case that the terms that are connected by hyphens are considered to be composed terms. This pushed certain authors to use abbreviations in order to shorten the texts that contained biologic entities made up of several terms. The use of abbreviations is a very common practice in the biomedical literature and the amount of acronyms is constantly growing. Chang and Schutze [4] consider that the number of abbreviations present in the *MEDLINE* database increases with an average of 400000 per year. In fact, it is estimated that a new acronym is added for every 5-10 abstracts [4]. Another consequence of the composition of entities made up of several terms concerns the usage of gene names with the permutation, insertion or removal of certain constituting words. For example, the entity *focal adhesion associated kinas* can also be referred to as *focal adhesion kinase* [32]. In addition, certain genes' names can be used with upper or lower cases letters (NF-KB and NF-kb) or even contain special characters. The variation of terms can also occur due to the spelling based on pronunciation (*tumour* and *tumor*) or due to Latin variations (*oestrogen* and *estrogen*) [2, 4].

Terms ambiguity:

This occurs in cases where two or more entities can have the same name but refer to terms that belong to different fields. This happens in the case of the *Cdc2* gene that actually refers to two different genes (*budding* and *fission yeas*) [17]. For example, the gene AR is often used for referring to the following terms [32]:

[4] http://flybase.org

- Androgen Receptor
- AmphiRegulin
- Acyclic
- Retinoid
- Agonist-Receptor
- Adrenergic
- Receptor

Although these entities share the same abbreviation (AR), they actually refer to different genes; which underlines the ambiguous aspect of the biologic entities.

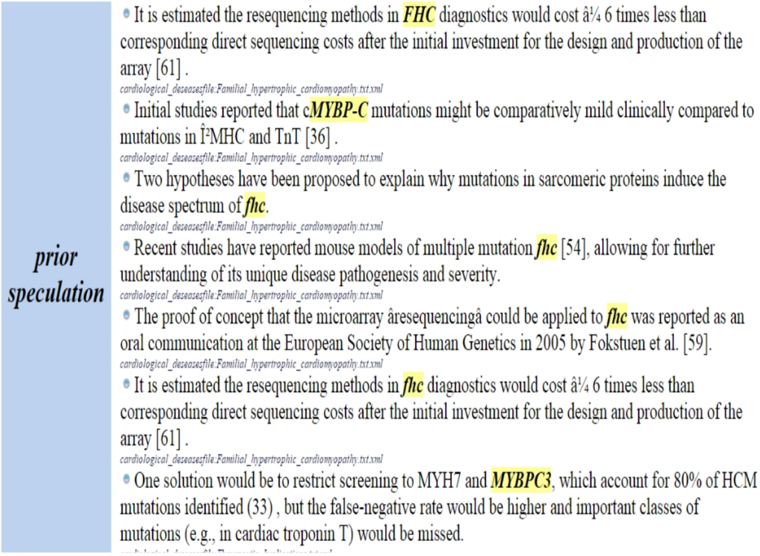
Introduction of new terms:
Another difficulty that concerns the research of information based on the biologic entities is the usage of new terms in the biomedical literature. This is due to the continuous development of research within the biomedical field that triggered a remarkable increase in the amount of new terms used for defining the biologic entities. Within this framework, [28] analyze a corpus of 52,845 abstracts of *MEDLINE*[5] in relation to the bakers yeast subject. The comparison of this corpus with another one made up of fulltext articles from chemistry journals that belong to the biomedical field showed the presence of new data.

Our goal consists in attempting to cover the majority of occurrences of terms that can be linked to a particular biological entity taking into account these possible variations. We propose in this context to use a process that enables the connection to various biomedical databases available on line (GPSDB[6] : Gene and Protein Synonym DataBase). It is a database accessible through an interface that extracts a list of possible variations of a particular entity in the text [29].This not only solves the problems associated with variations in the terms but also takes into account the new recently introduced in the literature entities. For example, a user can be interested in retrieving new speculations issued around the FHC gene. In this case, he enters the name of the gene to look for, a list of possible variations of this gene is displayed which enables him to select a set of search terms in the annotated sentences. In our example, the user selects from the list, in addition to the FHC gene, the gene MYBPC3. The resulted synthesized sheet is presented in Fig.(2.7).

2.6 Related work

The annotation models of scientific publications proposed in previous works are based on two approaches: Annotation models that classify information according to section titles and Annotation models that propose a fine-gained annotation categorization.

[5] http://www.nlm.nih.gov/bsd/pmresources.html
[6] http://www.pharmadm.com/biomint/org/



prior speculation

- It is estimated the resequencing methods in ***FHC*** diagnostics would cost â¼ 6 times less than corresponding direct sequencing costs after the initial investment for the design and production of the array [61] .
  cardiological_deseasesfile:Familial_hypertrophic_cardiomyopathy.txt.xml
- Initial studies reported that c***MYBP-C*** mutations might be comparatively mild clinically compared to mutations in Î²MHC and TnT [36] .
  cardiological_deseasesfile:Familial_hypertrophic_cardiomyopathy.txt.xml
- Two hypotheses have been proposed to explain why mutations in sarcomeric proteins induce the disease spectrum of ***fhc***.
  cardiological_deseasesfile:Familial_hypertrophic_cardiomyopathy.txt.xml
- Recent studies have reported mouse models of multiple mutation ***fhc*** [54], allowing for further understanding of its unique disease pathogenesis and severity.
  cardiological_deseasesfile:Familial_hypertrophic_cardiomyopathy.txt.xml
- The proof of concept that the microarray âresequencingâ could be applied to ***fhc*** was reported as an oral communication at the European Society of Human Genetics in 2005 by Fokstuen et al. [59].
  cardiological_deseasesfile:Familial_hypertrophic_cardiomyopathy.txt.xml
- It is estimated the resequencing methods in ***fhc*** diagnostics would cost â¼ 6 times less than corresponding direct sequencing costs after the initial investment for the design and production of the array [61] .
  cardiological_deseasesfile:Familial_hypertrophic_cardiomyopathy.txt.xml
- One solution would be to restrict screening to MYH7 and ***MYBPC3***, which account for 80% of HCM mutations identified (33) , but the false-negative rate would be higher and important classes of mutations (e.g., in cardiac troponin T) would be missed.

Figure 2.7: Visualization of a Multi-documents synthesized sheet related to the FHC gene entity.

In the first approach the annotation models aim to annotate texts according to titles of sections that appear frequently in scientific papers such as “*Introduction*”, “*Method*”, “*Result*“ and “*Conclusion*”. Indeed, Hirohata et al. [16] classify abstract sentences into: “*Objectif*“, “*Method*“, “*Result*” and “*Conclusion*“. In order to classify automatically these sentences they used Conditional Random Fields (CRFs) and obtained 95.5% for sentences and 68.8 % for abstracts.

The annotation model proposed in [22] is composed of “*Introduction*”, “*Method*”, “*Result*” and “*Conclusion*”. Authors use for this task Hidden Markov Model (HMM) and obtain an F-score of 88.5%, 84.3%, 89.8% and 89.7% for the categories “*Introduction*”, “*Method*”, “*Result*” and “*Conclusion*”, respectively.

Shimbo et al. [31] use SVM to classify sentences in Medline abstracts into “*Objectif*”, “*Method*”, “*Result*” and “*Conclusion*”. The classification results are 91.9% of precision for sentences and 51.2% of precision for abstracts.

The system proposed by Yamamoto and Takagi [38] aims to classify sentences according to five categories: “*Context*”, “*Objectif*”, *“Method*”, “*Result*” and *“Conclusion*” using SVM. Results obtained by this method are: 68.9%, 63%, 83.6%, 87.2% and 89.8% for respectively the “*Context*”, “*Objectif*”, “*Method*”, “*Result*” and “*Conclusion*”.

The semantic map that we propose belongs to the second approach: Annotation models that propose a fine-gained annotation categorization. In this context, Teufel and Moens [34] 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*”) where the annotation process is performed by Bayesians classifiers. The results of this method are between 26% as minimum F-measure (obtained for the "*Contrast*" category) and 86% as maximum F-measure (obtained for the "*Other*" category). This model was then improved in [35] where authors introduce the AZ II model that aims to annotate information related to publications in the biomedical domain. The semantic categories of the AZ annotation model [34] are also used by Mizuta et al. [25] who improved the model by proposing their own seven categories ("*Background*", "*Problem*", "*Outline*", "*Textual*", "*Own*", "*Connexion*", and "*Difference*"). The model implementation is realized using SVM and Bayesian classifiers [26]. The evaluation results achieve an F-Score of 70%.

deWaard et Pander Maat [6] studied the ABCDE (Annotation, Background, Contribution, Discussion and Entities) theoretical structure of scientific papers and identified seven types of epistemic segments: "*Fact*"," *Hypothesis*", "*Implication*", "*Goal*", "*Method*", "*Result*" and "*Problem*".

In [20], the author developed the CoreSC (Coresc Scientifique Concepts) annotation model that contains the following categories: "*Hypothesis*", "*Motivation*", "*Background*", "*Objectif*", "*Object*", "*Experience*", "*Model*", "*Method*", "*Observation*", "*Result*" and "*Conclusion*". The automatic recognition of the CoreSC categories is performed using machine learning tools and were evaluated on a corpus of 265 full paper linked to the biochemistry and chemistry domain. The classification results are between 18% for the "*Motivation*" category and 76% for the "*Experiment*" category. Nawaz et al. [27] and Thompson et al. [36] propose a multi-dimensional model in order to annotate biological events according to various dimensions: Knowledge Type ("*Demonstrative*", "*Deductive*", "*Sensorial*" and "*Speculation*"), Certainty Level ("*Absolute*", "*high*", "*moderate*" and "*low*") and Point of view (to indicate if the declaration is based on the author point of view or if it deals with reported information).

2.7 Discussion

The semantic map that we propose is based on an automatic categorization of the information situated among the works of the second approach. The comparison between this semantic map and one of the models of the approach that is based on the titles of the sections (that of [16]) shows that there are common categories *"Method"* and *"Result"*. However, some of these common categories do not cover the same information. For example, the category *"Result"* belonging to Hirohata's model [16] can in fact correspond to the sub-categories *"New Result"*, *"New Speculation"* or even new deduction from our semantic map.

It should be also noted that most of the annotation models from the first approach aim to analyze abstract papers. In fact, many authors consider that the information found in the abstracts is more important than that of the articles' main bodies. Demner-Fushman and Lin [8] suggest that, the information found in these abstracts offer sufficient indicators for the identification of papers that could potentially be interesting for the readers. However, Cohen and Hunter [5] carried out a comparative study analyzing the information existing in the full texts and that presented in the ab-

Table 2.7: Statistics of non-textual elements and their comments.

	Paper 1	**Paper 2**	**Paper 3**
Non textual elements	8	7	5
Sentences that comment non textual elements	48	12	21
Number of sentences that comment non textual elements annotated by the semantic map	21	3	12
Total number of sentences	414	324	292

stracts. This study showed that the sentences of the abstracts are less complex than those found in the complete versions of the papers (full text) and that, consequently, the textual search tools are more efficient for processing the abstracts of the articles than for processing full texts. Nonetheless, the relevance of the information is better perceived in the entire texts than in the abstracts. By contrast to the processing of the first approach, we consider it is preferable to have methods able to perform an automatic search of full texts in order to better identify the important information that were not kept in the abstracts.

Also, the analysis of the information presented in the abstract papers does not enable the processing of the non-textual data (figures, images, tables, graphics, etc.) that are integrated in the body of the text. We believe that this latter task is crucial in the process of construction of synthesized sheets. Contrary to the majority of previous works, the semantic map that we presented allows the identification of non-textual elements that are present in the scientific publications but also those in the associated comments. The latter ones can be related to one or more categories of the semantic map (for instance *"Result"*, *"Method"*).

In order to analyze the importance of this information, we proceeded with a manual annotation of three articles (chosen randomly) from different scientific journals. The results of this analysis are detailed in the Table 2.7. Broadly speaking, the analysis of these publications shows that there is an important amount of non-textual elements that can, for example, describe results or provide details on a certain methodology used. In Fig.(2.8), the corresponding legend and its content indicate that the author illustrated his results by means of a non-textual element.

A considerable amount of sentences that belong to one of the categories of the semantic map (result, method) refers to non-textual elements. These provide a more detailed description of the content of the information. For instance, for the first paper (Table 2.7) approximately 44% of the sentences that belong to a category of our semantic map refer to non-textual elements. The following sentence (that is in fact comments of non-textual elements) belongs to the category "*Result*". The research of a piece of information that focuses solely on the textual content is therefore insufficient. Consequently, the use of the association between non-textual elements and their relating comments in the process of research of data proves itself necessary.

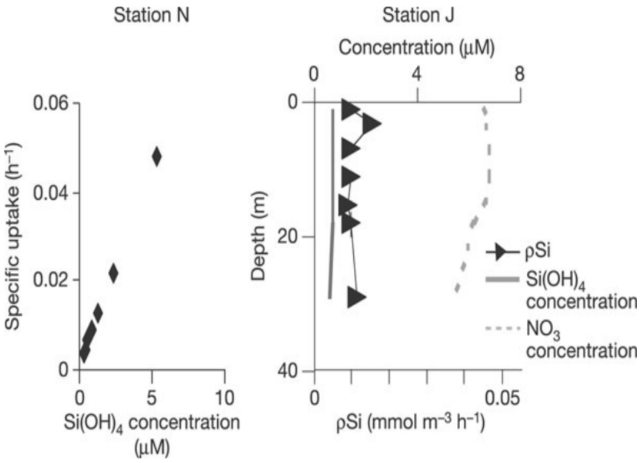


Figure 3 The results of Si(OH)4 uptake incubation experiments

Figure 2.8: Example of a non-textual element.

(35) "*Results showed that incubation with increasing concentrations of SCR7 inhibitedthe formation of multimers at 200 µM and above, unlike SCR5* ***(Figures 2B and 2C)***."

The analysis of the annotation models belonging to the second approach reveals that they have more common categories. For example, the category "*Topic announcement*" of our semantic map corresponds to the category "*Background*" from the annotation model of Waard and Pander Maat [6]. However, certain annotation models recommend subcategories that can depend on different criteria. For instance, the annotation model CoreSC [20] presents a categorization that allows the identification of experimental methods.

According to us, the process of description and search for the realized experiences, as well as their comparison with other works, is based not only on the importance of the interpretations provided by the authors, but also on the certainty level of these interpretations. This concept of certainty level is taken into account in the elaboration of the semantic map that we propose. The latter enables the classification of certain categories by level of reliability from the highest level to the lowest one: "*Result*", "*Conclusion*", "*Deduction*", "*Speculation*", "*Absence of knowledge*". These categories are very similar to those introduced by Thompson et al. [36] however; they are linked to the notion of types of knowledge separated from the notion of degree of reliability. We will not make this distinction and therefore we will maintain a correspondence between the degree of reliability and the types of knowledge because we consider that, in these scientific publications, these notions are connected in an intrinsic manner.

2.8 Conclusions

Our contribution is to present a general method that is based on the automatic and semantic annotation in order to extract information from scientific publications. The textual mining task is realized according to semantic categories such as the identification of new results or the identification of new hypothesis. Although the linguistic resources of our system are domain independent, we envisage evaluating the system annotations on a corpus related to various domains in order to confirm our annotation methodology.

References

[1] M. Alrahabi. *EXCOM-2: plate-forme d'annotation automatique de catégories sémantiques: Applications a la catégorisation des citations en français et en arabe*. PhD thesis, Université Paris-Sorbonne, France, 2010.

[2] S. Ananiadou and J. Mcnaught. *Text Mining for Biology and Biomedicine*. Artech House,Inc. Norwood, MA, USA, 2005.

[3] S. Ananiadou, D. Sullivan, W. Black, G.-A. Levow, J.J. Gillespie, C. Mao, S. Pyysalo, B. Kolluri, J. Tsujii, and B. Sobral. Named entity recognition for bacterial type IV secretion systems. *PLoS One*, 6(3):e14780, 2011.

[4] J. Chang and H. Schutze. *Text Mining for Biology and Biomedicine*, chapter Abbreviations in biomedical text, pages 99–119. Artech House,Inc. Norwood, MA, USA, 2005.

[5] K.B. Cohen and L. Hunter. Getting started in text mining. *PLoS Computational Biology*, 4(1):e20, 2008.

[6] A. de Waard and H. Pander Maat. Categorizing epistemic segment types in biology research articles. In *Workshop on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS)*, 2009.

[7] A. de Waard, S. Buckingham Shum, A. Carusi, J. Park, M. Samwald, and Á. Sándor. Hypotheses, Evidence and Relationships: the HypER approach for representing scientific knowledge claims. In *Workshop on Semantic Web Applications in Scientific Discourse (SWASD)*, 2009.

[8] D. Demmer-Fushan and J. Lin. Knowledge extraction for clinical question answering: preliminary results. In *Workshop on Question Answering in Restricted Domains (AAAI)*, pages 9–13, 2005.

[9] J. Desclés, M. Alrahabi, and J.-P. Desclés. *Human Language Technology*, chapter BioExcom: detection and categorization of speculative sentences in biomedical literature, pages 478–489. Springer Berlin Heidelberg, 2011.

[10] J. Desclés and O. Makkaouiand J.-P. Desclés. Towards automatic thematic sheets based on discursive categories in biomedical literature. In *International Conference on Web Intelligence, Mining and Semantics, page 32. ACM.*, 2011. Article No.32.

[11] J. Desclés, O. Makkaoui, and T. Hacène. Automatic annotation of speculation in biomedical texts : new perspectives and large-scale evaluation. In *Workshop on Negation and Speculation in Natural Language Processing*, pages 32–40, 2010.

[12] J.-P. Desclés. Systèmes dexploration contextuelle. *Co-texte et calcul du sens*, pages 215–232, 1997.

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

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

[15] F.A.P. Harmsze. *A modular structure for scientific articles in an electronic environment*. PhD thesis, FNWI: Van der Waals-Zeeman Institute, Amsterdam, Netherlands, 2000.

[16] K. Hirohata, N. Okazaki, S. Ananiadou, M. Ishizuka, and M. I. Biocentre. Identifying sections in scientific abstracts using conditional random fields. In *3rd International Joint Conference on Natural Language Processing (IJCNLP)*, pages 381–388, 2008.

[17] L.J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7(2):119–129, 2006.

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

[19] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsuji. Genia corpus a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):180–182, 2003.

[20] M. Liakata. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Workshop on Negation and Speculation in Natural Language Processing*, pages 1–4, 2010.

[21] M. Light, X.Y. Qiu, and P. Srinivasan. The language of bioscience: facts, speculations, and statements in between. In *Workshop on Linking Biological Literature Ontologies and Database*, pages 17–24, 2004.

[22] J. Lin, D. Karakos, D. Demner-Fushman, and S. Khudanpur. Generative content models for structural analysis of medical abstracts. In *HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 65–72, 2006.

[23] O. Makkaoui, J. Desclés, and J.-P. Desclès. Evaluation and performance improvement of the BioExcom system for the automatic detection of speculation in biomedical texts. In *3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*, 2012.

[24] B. Medlock and T. Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2007.

[25] Y. Mizuta, A. Korhonen, T. Mullen, and N. Collier. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487, 2006.

[26] T. Mullen, Y. Mizuta, and N. Collier. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *ACM SIGKDD Explorations Newsletter - Natural language processing and text mining*, 7(1):52–58, 2005.

[27] R. Nawaz, P. Thompson, J. McNaught, and S. Ananiadou. Meta-knowledge annotation of bio-events. In *LREC*, pages 2498–2505, 2010.

[28] G. Nenadic, I. Spasic, and S. Ananidou. Mining biomedical abstracts: what's in a term? In *1st International Joint Conference on Natural Language Processing*, pages 797–806, 2004.

[29] V. Pillet, M. Zehder, A.K. Seewald, A.-L. Veuthey, and J. Petrak. GPSDB: a new database for synonyms expansion of gene and protein names. *Bioinformatics*, 21(8):1743–1744, 2005.

[30] F. Le Priol. Automatic annotation of images, pictures or videos comments for text mining guided by no textual data. In *21st International conference Florida Artificial Intelligence Research Society (FLAIRS)*, pages 494–499, 2008.

[31] M. Shimbo, T. Yamasaki, and Y. Matsumoto. Using sectioning information for text retrieval: a case study with the medline abstracts. In *2nd International Workshop on Active Mining (AM)*, pages 32–41, 2003.

[32] I. Spasic, S. Ananiadou, J. Mcnaught, and A. Kumar. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251, 2005.

[33] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, 2008.

[34] S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[35] S. Teufel, A. Siddharthan, and C. Batchelor. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Conference on EmpiricalMethods in Natural Language Processing*, volume 3, pages 1493–1502, 2009.

[36] P. Thompson, R. Nawaz, J. McNaught, and S. Ananiadou. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12(1):393, 2011.

[37] J.D. Wren, J.T. Chang, J. Pustejovsky, E. Adar, H.R. Garner, and R.B. Altma. Biomedical term mapping databases. *Nucleic Acids Research*, 33(suppl 1):D289–D293, 2005.

[38] Y. Yamamoto and T. Takagi. A sentence classification system for multi biomedical literature summarization. In *21st International Conference on Data Engineering Workshops (ICDE)*, pages 1163–1163, 2005.