

An Automated Annotation Process for the SciDocAnnot Scientific Document Model

Hélène de Ribaupierre^{1,2} and Gilles Falquet¹

¹ CUI, University of Geneva, 7, route de Drize, CH 1227 Carouge, Switzerland

² Department of Computer Science, University of Oxford, UK

{Helene.deribaupierre, Gilles.falquet}@unige.ch

Abstract. Answering precise and complex queries on a corpus of scientific documents requires a precise modelling of the document contents. In particular, each document element must be characterised by its discourse type (hypothesis, definition, result, method, etc.). In this paper we present a scientific document model (SciAnnotDoc) that takes into account the discourse types. Then we show that an automated process can effectively analyse documents to determine the discourse type of each element. The process, based on syntactic rules (patterns), has been evaluated in terms of precision and recall on a representative corpus of more than 1000 articles in Gender studies. It has been used to create a SciDocAnnot representation of the corpus on top of which we built a faceted search interface. Experiments with users show that searching with this interface clearly outperforms standard keyword search for complex queries.

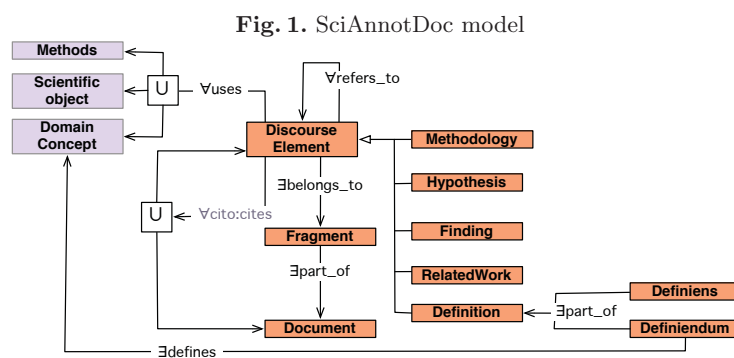
1 Introduction

One of the challenges, today, for Information Retrieval System for Scientific Document is to fulfil the information needs of scientists. For scientists, being aware of others' work and publications in the world is a crucial task, not only to stay competitive but also to build their work upon already proven knowledge. In 1945, Bush [2] already argued that too many publications can be a problem because the information contained in these publications cannot reach other scientists. Bush expounded his argument using the example of Mendel's laws of genetics. These laws were lost to the world for a generation because the Mendel's publication did not reach the few people who were capable of understanding and extending this concept. Today this problem is even more important with the exponential growth of literature in all domains (e.g., Medline has a growth rate of 0.5 million items per year [8]). Today's IR systems are not able to answer precisely to queries such as "find all the definition of the term X" or "find all the findings that analyse why the number of women in academics falls more sharply than the number of men after their first child, using qualitative and quantitative methodologies". These systems are in general using only the metadata of the documents to index them (title, author(s), keywords, abstract, etc.), but to obtain systems that answer to such precise queries, we need to have very precise

semantic annotation of the entire documents. In [13],[15], we have proposed a new annotation model for scientific document (SciAnnotDoc annotation model). The SciAnnotDoc (see Figure 1) annotation model is a generic model for scientific documents. This model can be decomposed in four different dimensions or facets:

1. Conceptual dimension: Ontologies or controlled vocabularies that describe scientific terms (the SciDeo ontology) or concepts used in the document (conceptual indexing)
2. Meta-data dimension: description of meta-data document information (bibliographic notice)
3. Rhetorical or discursive dimension: description of the discursive role played by each document element
4. Relationships dimension: description of the citations and relationships between documents

The third facet is extremely important when considering precise scientific queries and is decomposed into five discourse elements types that are: *findings*, *hypothesis*, *definition*, *methodology* and *related work*. We retained these five discourses elements after analysing the result of a survey and interviews made with scientists in different field of research to determine what and how scientists are searching and reading scientific documents [12],[14]. The SciAnnotDoc model is implemented in OWL. The ontology contains 69 classes, 137 object properties and 13 datatype properties (counting those imported from CiTO³[18]). The model also integrate ontologies that help in the annotation process (the violet ontologies), and that are given more information about the content such as domain concept, scientific object or methods names contained in the different discourse element.



³ CiTO is used to describe the different type of citation or reference between the documents or discourse element

In this paper, we present the automatic annotation processes we used to annotate scientific documents using the SciAnnotDoc model. The process is based on natural language processing (NLP) techniques.

To evaluate the annotation process, we used a corpus in gender studies. We chose this domain because it consists of very heterogeneous written documents ranging from highly empirical studies to "philosophical" texts, and these documents are less structured than in other fields of research (i.e. medicine, biomedicine, physic, etc.) and rarely use the IMRaD model (introduction, methods, results and discussion). This corpus is therefore more difficult to annotate than a corpus of medical documents, which is precisely the kind of challenge we were looking for. We argue that if our annotation process can be applied to such a heterogeneous corpus, it should also apply to other, more homogeneous types of papers. Therefore, the annotation process should be generalisable to other domains.

In the literature, there are three types of methods for the automatic annotation or classification of scientific text documents. The first type is rule based. This type of system is based on the detection of general patterns in sentences [4],[20],[17]. Several systems are freely available, such as XIP⁴, EXCOM⁵, and GATE⁶. The second method is based on machine learning and requires a training corpus, such as the systems described by [10],[19],[16],[6]. Several classifiers are available, such as the Stanford Classifier⁷, Weka⁸, and Mahout⁹, based on different algorithms (Decision Trees, Neural Networks, Naïve Bayes, etc.,¹⁰). The third method is a hybrid between the two aforementioned systems [9].

In this work, we opted for a rule-based system because we did not have a training corpus and because documents in the human sciences are generally less formalised than in other domains, it may be difficult to have sufficient features with which to distinguish the different categories. Between the several free semantic annotation tools existing, we choose to use GATE, because it is used by a very large community and several plug-ins are available.

2 Annotation Implementation

The annotation processes transforms each sentence into a discourse element (or if it is not one of the five discourse element into a non defined discourse element) and a paragraph into a fragment. Each fragment contains one to many discourses elements and each sentence can be attributed to one or many discourse elements (e.g. a sentence that describe a definition can be also a sentence that describe a finding). The following sentence will be annotated as a definition and a finding.

⁴ <https://open.xerox.com/Services/XIPParser>

⁵ <http://www.excom.fr>

⁶ <http://gate.ac.uk>

⁷ <http://nlp.stanford.edu/software/classifier.shtml>

⁸ <http://www.cs.waikato.ac.nz/ml/weka/>

⁹ <https://mahout.apache.org/users/basics/algorithms.html>

¹⁰ see [5] for a complete review of the different algorithms

”We find, for example, that when we use a definition of time poverty that relies in part on the fact that an individual belongs to a household that is consumption poor, time poverty affects women even more, and is especially prevalent in rural areas, where infrastructure needs are highest.”[1]

The discourse element *related work* is a special case, it will always be defined at first as one of the four other discourses elements and defined as a related work thereafter. The reason for this kind of annotation is the results of the analyses we made from the interviews. Scientists sometimes are looking for a finding, a definition, a methodology or a hypothesis but the attribution of the document to an author is not in their priority, it is only later that they might be interested to know who the author of the document are or the referenced sentences. For example, the following sentence is a finding and in a second time a related work, as this sentence refer to another works.

”The results of the companion study (Correll 2001), and less directly the results of Eccles (1994; Eccles et al. 1999), provide evidence that is consistent with the main causal hypothesis that cultural beliefs about gender differentially bias men and women’s self-assessments of task competence”.[3]

In a first step, to discover and analyse the syntactic patterns of the discourse elements, we manually extracted sentences that corresponded to the different discourse elements from scientific documents in two areas of research: computer science and gender studies. In a second step, we uploaded these sentences in GATE and ran a pipeline composed of components included in ANNIE (ANNIE Tokeniser component, the ANNIE sentence splitter component and the ANNIE part-of-speech component) and obtained the syntactic structures of these sentences. The aim of this analyse was to build the JAPE rules for detecting the different discourses elements. The methodology used to create the rules was the following. First, we started to look at the syntactic structure produced by the ANNIE output for each of the different sentences. The following example (see table 1) describes the entire tag’s sequence obtained by ANNIE on the following definition of the term ”gender” (for space reason, we didn’t write down all the sequence¹¹).

”On this usage, gender is typically thought to refer to personality traits and behavior in distinction from the body”.[7]

For each tag’s sequence, we simplified those rules, reduced them and merged some of them, to obtain more generic rules able, not only to catch the very specific syntactic pattern, but also to catch the variation of the pattern. We also

¹¹ the definition of the part-of-speech tags can be found at <http://gate.ac.uk/sale/tao/splitap7.html#x39-789000G>

Table 1. ANNIE tag's sequence

On	this	usage	,	gender	is	typically	thought	to	refer	to	personality	traits	and
IN	DT	NN	,	NN	VBZ	RB	VBN	TO	VB	TO	NN	NNS	CC

relaxed the rules and used some unspecified token (see table 2). To increase the precision, we added typical terms that appear in each type of discourse element. For the definition example, instead of using the tag VBZ¹² that could be too generic, we used a macro that was the inflection of the verb *be* and *have* in the singular and plural form. We also used a macro that defined the different inflection of the verb *refer to*. With this simplification and relaxation, we can now annotate sentences such as those shown in the table 2.

Table 2. Definition sentences and JAPE rules (simplified)

gender	is	typically thought to	refer to
gender	has	become used to	refer to
gender	was	a term used to	refer to
NN	TO_BE_HAVE(macro)	Token[2,5]	REFER(macro)

We uploaded the domain concept ontology to help to define more precise rules. For example, to detect a definition such as "It follows then that gender is the social organization of sexual difference" [7]; we created a rule that was searching for a concept defined in the domain ontology followed at a short distance by the declension of the verb *be*.

```
((({Lookup.classURI==".../genderStudies.owl#concept"})
(PUNCT)?
(VERBE_BE))
```

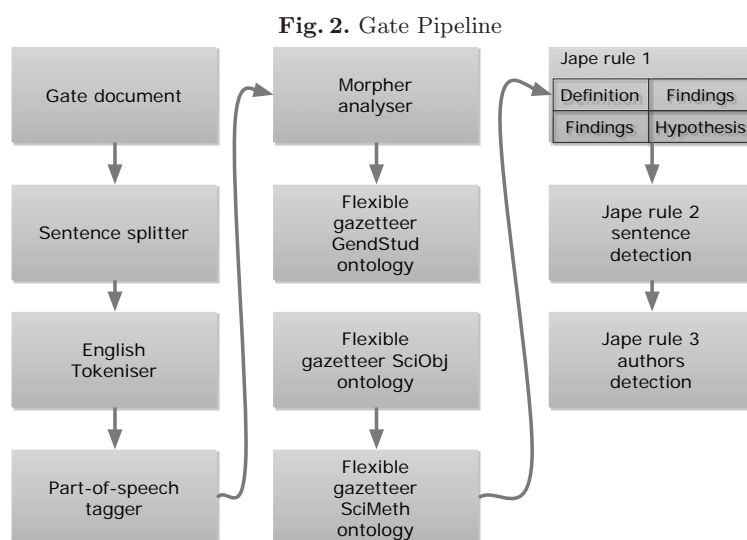
To be able to use the different ontologies, we used the ontologies plug-in contained in GATE¹³. We imported the different ontologies that we created to help the annotation process: the gender studies ontology (GenStud), the scientific ontology (SciObj) and the methodology ontology (SciMeth). The ontologies were used not only for the JAPE rules but also to annotate the concepts in the text. With this methodology we defined 20 rules for findings, 34 rules for definitions, 11 rules for hypothesis and 19 rules for methodologies and 10 rules for the referenced sentences.

We automatically annotated 1,400 documents in English from various journals in gender and sociological studies. The first step consisted of transforming a PDF file into a raw text. PDF is the most frequently used format to publish

¹² 3rd person singular present

¹³ Ontology OWLIM2, OntoRoot Gazetteer

scientific documents, but it is not the most convenient one to transform into raw text. The Java program implemented to transform the PDF into raw text used the PDFbox¹⁴ API and regular expressions to clean the raw text. Second, we applied the GATE pipeline (see figure 2). The output given by GATE is a XML file.



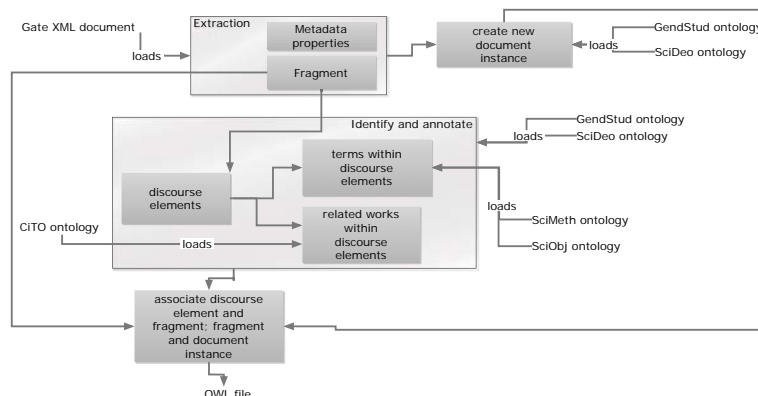
Third, we implemented a Java application (see figure 3) using the OWL API to transform the GATE's XML files into an RDF representation of the text. Each XML tag corresponding to concept or object properties in the ontologies were transformed. The sentences that did not contain one of the four discourse elements (definition, hypothesis, finding or methodology) were annotated with the tag `<NonDefinedDE>`, allowing for the annotation of each sentence of the entire document, even those not assigned to discourse elements. Each discourse element that contained a tag with `<AuthorRefer>` were defined as a related work. The different RDF representations created by the Java application were loaded into an RDF triple store. We chose Allegrograph¹⁵ because it supports *RDFS* ++ reasoning in addition to SPARQL query execution.

The table 3 presents the distribution of the discourse elements by journal. We can observe that the number of referenced documents is greater than the number of related works. This result is observed because authors generally refer to several documents simultaneously rather than to a single one. We can also observe that the most fund discourse element is the finding, followed by methodology, followed

¹⁴ <https://pdfbox.apache.org/>

¹⁵ <http://franz.com/agraph/allegrograph/>

Fig. 3. Annotation algorithm model



by hypothesis and definition. This distribution seems to follow the hypothesis that scientist communicate more their finding than everything else, even in field of research such as sociology or gender Studies. And it is also the most researched discourse element that from the survey and the interviews scientists are looking for [12],[14].

Table 3. Annotated corpus statistics by Discourse elements

Journal Name	Def.	Find.	Hypo.	Meth.	Related Work	Referenced documents
Gender and Society	745	2945	1021	1742	986	4855
Feminist Studies	2201	4091	2545	3660	177	5377
Gender Issues	280	1126	414	611	267	1566
Signs	789	1566	712	1221	516	3129
American Historical Review	97	219	87	170	15	440
American Journal Of Sociology	1776	10160	4316	6742	2907	13323
Feminist economist	1381	6940	2025	4169	2288	9600
Total	7269	27047	11120	18315	7156	38290

To test the quality of the patterns, we uploaded 555 manually annotated sentences that constitute our gold standard into GATE and processed through the same pipeline (see Figure 2). We did not use any of the sentences analysed to create the JAPE rules to construct the gold standard to avoid bias. We performed measurements of precision and recall on these sentences (see Table 4). The results indicated good precision but a lower recall. One of the reasons of the lower recall could be that the JAPE rules are very conservative.

Table 4. Precision/recall values

Discourse element type	No of sentences	Precision	Recall	F1.0s
Findings	168	0.82	0.39	0.53
Hypothesis	104	0.62	0.29	0.39
Definitions	111	0.80	0.32	0.46
Methodology	172	0.83	0.46	0.59

3 User evaluation on complex queries

We conducted user evaluations to check how the annotation system and the SciDocAnnot model compare to standard keyword search. We implemented two interactive search interfaces: a classic keywords based search (with a TF*IDF based weighting scheme) and a faceted interface (FSAD) based on our model (facets correspond to the types of discourse elements). Both systems are indexing and querying at the sentence level (instead of the usual document level). The first tests we conducted with 8 users (scientists, 4 in gender studies and 50% women with an average age of 38 years old). Scientists had to perform 3 tasks with only one of the system (see below). The design of the experiment was based on a Latin square rotation of tasks to control for a possible learning effect of the interface on the participants.

task 1 Find all the definitions of the term "feminism".

task 2 Show all findings of studies that have addressed the issue of gender inequality in academia.

task 3 Show all findings of studies that have addressed the issue of gender equality in terms of salary.

We gave them a small tutorial on how the system works, but didn't give more exact instruction on how to search. The participants, who decided whether they had obtained enough information on the given subject, determined the end of a task. They have to perform the task and complete 4 different questionnaire (1 socio-demographic, 1 after each task and a final at the end of the evaluation, for more precision about the questionnaires see [11]). The questionnaire after each task contained 10 questions and the last questionnaire 11 questions; most of the questions were using a Likert scale. The evaluation was performed in French. The questionnaires were conducted on LimeSurvey. We computed the average response for the three tasks and we tested the difference between the participants who had to evaluate the FSAD versus the keyword search, using an analysis of the variance (Anova) tests. Because of the lack of space, in this paper, we will only present a part of the evaluation.

The first question (Do you think the set of results was relevant to the task? 1=not useful, 5=useful) was about the relevance of the set of results,. We didn't observe any significant difference between the two groups of user, both found that the set of answer was useful (FSAD M=4.0; keywords M=3.75). But in

the second question (Do you think the number of results was too large to be useful? 1 = totally unusable, 5 = usable), about the irrelevance of the question, the keywords-search group (M=2.75) find the irrelevance of the set of answer more important than the FSAD group (M=4.5), and a significant difference was observed ($p < 0.05$) between the two groups. We also asked the same question, but instead to have to answer with a Likert scale, users have to answer with a scale in percent (How many elements correspond to your request? 1 = 0-5%, 2= 6-15%, 3= 16-30%, 4=31-50%, 5=51-75%, 6=76-90%, 7,+90%). Again we didn't find a significant difference in this question between the group (FSAD M=5.0; keywords M=4.25), but for the second question, we again find a significant difference between the two group (FSAD M=1.5; keywords M=3.75; $p < 0.05$). When we asked the users for the level of satisfaction they experimented with the set of results (Did you obtain satisfactory results for each query you made? 1 = not at all satisfied, 5 = very satisfied), the difference between the two groups is not significant (FSAD M=3.83; Keywords M=3.41). The next question was about the level of satisfaction overall the set of results for the whole task (Are you satisfied with the overall results provided? 1 = not at all satisfied, 5 = completely satisfied). Most of the users did more than one query by task. The participants who used the keyword search interface seemed to be less satisfied with the overall results than the participants who used the FSAD, but the difference was not significant (FSAD M=4.16, Keywords M=3.41). We also ask the user about their level of frustration to the set of results (Are you frustrated by the set(s) of results provided? 1 = totally frustrated, 5 = not at all frustrated). The FSAD group seemed to be less frustrated with the set of results than the keyword search interface group, but the difference was not significant (FSAD M=3.16; keywords M=4.25).

Aside from the user evaluation, we also performed a precision and recall evaluation for the first task. For the FSAD system, when the user choose the facet definition and type the keyword feminism, the system send a set of 148 answers and 90 were relevant, the precision was 0.61. For the keywords-search system, the set of result was the combination of the term "define AND feminism" and "definition AND feminism" (this combination of term was the one the most used by users for the task 1), the system sent a set of 29 answers of which 24 were relevant), the precision was 0.82.

For the recall, as we didn't know the number of definitions contained in the corpus, we simply observed that the ratio between FSAD and the keyword search is 3.77. In other words, the FSAD system was able to find 3.77 times more definitions of the term "gender" than the keywords-search. However even if the precision is slightly lower in the FSAD system than in the keywords-search system, the FSAD system has a considerably higher recall than the keyword search system.

4 Conclusion

The aim of this work is to propose an approach to help scientists to find the documents they need for their works. As presented in the introduction, it is important for scientists to have a search engine that is able to answer to precise questions such as "retrieve all the findings that women have a tendency to drop their academic carrier after their first child more than men, using qualitative and quantitative methodologies". In this case, knowing or indexing only the metadata is not enough, and annotation about the content of the full text such as the discourse element, the references to other documents and the concept is crucial. In this paper, we have proposed an approach to automatically annotated PDF document with the SciAnnotDoc model. The evaluation of the annotation show not only that the model is realistic because it is amenable to automatic production (many previously proposed annotation models have never been used in practice because they require a manual annotation, see [11], for a more complete review of the different systems and model), but also that the precision is good.

To improve the recall index, a solution could be to create more JAPE rules. However, introducing a larger number of rules might also increase the risk of adding noise to the annotation. Another solution could be to test if some hybrid approach mixing a rules-based approach and a machine-learning approach may improve precision and recall. Another solution is to ask experts not to classify the sentence in categories, but to confirm the type of categories a sentence is already classified into. By using this kind of methodology, we can improve and enlarge a training corpus that we could use to improve the precision and recall of the actual annotation process.

The evaluation with users shows that despite these inaccuracies and a small sample, we were able to build a query system that already outperforms keyword search in many cases, especially in the case where the recall is very important. Google allow to query a term for the definition with "define" + the term. In Google the set of answer is extracted from glossaries, dictionaries and Wikipedia for the first ranked answer, and for the next answers the system seems to work by looking at the pattern "define" + term. For scientists this is not enough, first because of the source of the information is not accurate enough and second because of the lack of answers. For Google Scholar, scientists make the assumption that the sources are more accurate because the IRs is indexing scientific documents. The system query the index with the pattern "define" AND "feminism", ignoring all the other definition that use some other sentence construction than "... define feminism...". And as we have shown above, the number of definition of the term found with this pattern is from a very long shot not enough, especially for scientists. By consequence, when the task is to find a definition and the user need a very high recall, Google or Google Scholar are not performing well. One of the difficulty we have to deal in the evaluation was the lack of a good evaluation corpus, helping us to calculate the precision and the recall of the system. This problem is very often mentioned in the literature, conference and workshops. We hope that in the future, with the different evaluation campaign that was created these last years, this recurrent problem should diminish.

The user evaluation also shown that the user seems to be less frustrated by the FSAD system than the keywords-search and seems to find that the level of irrelevance in the set of results is less important in FSAD than in keywords-search. Some of the results could be non significant because of the size of the sample that is a little bit too small.

In the case of very precise query such as the task 2 or 3 we still have to analyse the precision and recall of our system. We also want to compare the result with some today IR's, but we can hypothesis that in contrary of the first task, it will be the precision that will be missing because it is not searching at the sentence level. The reason is that Google and similar system are indexing the text by the terms they find in the metadata (title, abstract, keywords) and sometimes by the terms contained in the entire document, but they don't take into account the context of the term, or even the distance between the terms. For example, in the task 2, users will certainly types as keywords "academic" or "university" and "gender inequality", but the problem is that those terms could appear everywhere in the text, even in the references, and document that have for example a reference that was publish in the Oxford University Press and contains in an other part of the text "gender inequality" could appear in the top ranked answers.

In the future, we will conduct some additional usability testing and collect data to scientifically assess the quality of the system and to determine the influence of the precision/recall of the automated annotation process on the system performance. We will also conduct some experiment to analyse which kind of task is more demanding of a good precision, versus the one that need a good recall.

5 Acknowledgments

This work is supported by the Swiss National Fund (200020 138252)

References

1. Bardasi, E., Wodon, Q.: Working long hours and having no choice: time poverty in guinea. *Feminist Economics* 16(3), 45–78 (2010)
2. Bush, V.: As we may think. *The atlantic monthly* 176(1), 101–108 (1945)
3. Correll, S.J.: Constraints into preferences: Gender, status, and emerging career aspirations. *American Sociological Review* 69(1), 93–113 (2004), <http://asr.sagepub.com/content/69/1/93.abstract>
4. Groza, T., Handschuh, S., Bordea, G.: Towards automatic extraction of epistemic items from scientific publications. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. pp. 1341–1348. SAC '10, ACM, New York, NY, USA (2010), <http://doi.acm.org/10.1145/1774088.1774377>
5. Kotsiantis, S.B.: Supervised machine learning: A review of classification techniques. *Informatica* 31, 249–268 (2007)
6. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28(7), 991–1000 (2012)

7. Nicholson, L.: Interpreting gender. *Signs* 20(1), pp. 79–105 (1994), <http://www.jstor.org/stable/3174928>
8. Nováček, V., Groza, T., Handschuh, S., Decker, S.: CORAAL—Dive into publications, bathe in the knowledge. *Web Semantics: Science, Services and Agents on the World Wide Web* 8(2), 176–181 (2010)
9. Ou, S., Kho, C.S.G.: Aggregating search results for social science by extracting and organizing research concepts and relations. In: *SIGIR 2008 Workshop on Aggregated Search*, Singapour. Singapour (2008)
10. Park, D., Blake, C.: Identifying comparative claim sentences in full-text scientific articles. In: *50th Annual Meeting of the Association for Computational Linguistics*. pp. 1–9 (2012)
11. Ribaupierre, H.d.: Precise information retrieval in semantic scientific digital libraries. Ph.D. thesis, University of Geneva (2014)
12. Ribaupierre, H.d., Falquet, G.: New trends for reading scientific documents. In: *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*. pp. 19–24. *BooksOnline '11*, ACM, New York, NY, USA (2011), <http://doi.acm.org/10.1145/2064058.2064064>
13. Ribaupierre, H.d., Falquet, G.: A user-centric model to semantically annotate and retrieve scientific documents. In: *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*. pp. 21–24. ACM (2013)
14. Ribaupierre, H.d., Falquet, G.: Un modèle d'annotation sémantique centré sur les utilisateurs de documents scientifiques: cas d'utilisation dans les études genre. In: *IC-25èmes Journées francophones d'Ingénierie des Connaissances*. pp. 99–104 (2014)
15. Ribaupierre, H.d., Falquet, G.: User-centric design and evaluation of a semantic annotation model for scientific documents. In: *13th International Conference on Knowledge Management and Knowledge Technologies, I-KNOW '14*, Graz, Austria, September 16–19, 2014. pp. 1–6 (2014)
16. Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C., Veuthey, A.L.: Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics* 76(2–3), 195 – 200 (2007), <http://www.sciencedirect.com/science/article/pii/S1386505606001183>, *connecting Medical Informatics and Bio-Informatics - {MIE} 2005*
17. Sándor, Á., Vorndran, A.: Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries* pp. 36–44 (2009)
18. Shotton, D.: CiTO, the Citation Typing Ontology, and its use for annotation of reference lists and visualization of citation networks. *Bio-Ontologies 2009 Special Interest Group meeting at ISMB* (2009)
19. Teufel, S.: Argumentative zoning: Information extraction from scientific text. Unpublished PhD thesis, University of Edinburgh (1999)
20. Tutin, A., Grossmann, F., Falaise, A., Kraif, O.: Autour du projet scientext: étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. *Journées Linguistique de Corpus* 10, 12 (2009)