Semantic Facets for Scientific Information Retrieval

Iana Atanassova(✉) and Marc Bertin

CIRST, Université du Québec à Montréal, Succ. Centre-ville, B.P. 8888, Montreal, QC H3C 3P8, Canada
iana.atanassova@nlp-labs.org, bertin.marc@courrier.uqam.ca

Abstract. We present an Information Retrieval System for scientific publications that provides the possibility to filter results according to semantic facets. We use sentence-level semantic annotations that identify specific semantic relations in texts, such as methods, definitions, hypotheses, that correspond to common information needs related to scientific literature. The semantic annotations are obtained using a rule-based method that identifies linguistic clues organized into a linguistic ontology. The system is implemented using Solr Search Server and offers efficient search and navigation in scientific papers.

Keywords: Semantic annotation · Information retrieval · Faceted search · Semantic facets · Solr

1 Introduction

Today, the emergence of open science leads to the greater availability of scientific papers in full text. The ever larger volume of textual data provided fosters the development of new tools to explore the content of research papers. This problem has been studied from the point of view of the development of annotation frameworks for scientific papers [6,10]. Furthermore, the exploitation of this kind of annotations for information retrieval has been the object of many papers (e.g. [4,8]) and the extraction of key-phrases from scientific articles (see [11]) is a closely related subject.

In this paper, we describe a search engine that uses annotations related to a set of semantic categories as semantic facets in order to filter relevant information in scientific papers. The idea is to automatically identify specific discourse categories in the publications' content and make them directly accessible for the user to enhance text navigation and search. The goal of the development of semantic facets for information retrieval is to reduce the mental workload of users in the production of mental representations of documents in order to identify relevant information. This point of view has been discussed by Bertin and Atanassova [1].

V. Presutti et al. (Eds.): SemWebEval 2014, CCIS 475, pp. 108–113, 2014.
DOI: 10.1007/978-3-319-12024-9_14

Table 1. Dataset - PLOS journals

Journal	Number of articles	Number of sentences
PLOS Biology	2,965	426,522
PLOS Computational Biology	2,107	518,289
PLOS Genetics	2,560	566,323
PLOS Medicine	2,228	218,459
PLOS Neglected Tropical Diseases	1,366	217,861
PLOS Pathogens	2,354	514,751
PLOS ONE	33,782	6,080,566
Total	*47,362*	*8,542,771*

2 Semantic Annotation

For this study, we have processed research articles from seven journals, published by the Public Library of Science (PLOS) and available in Open Access. The articles are in the XML format, structured using the Journal Article Tag Suite (JATS), which provides the complete metadata and the full-text body of the articles. The sections and paragraphs in the text are represented as separate elements. We have processed the entire set of research articles of these journals up to September/October 2012. Table 1 presents the number of articles and sentences processed for each journal.

Metadata fields, such as titles, authors, abstract, journal and subject, are extracted from the XML documents. Additionally, we extract all the bibliographic data, i.e. the list of references in the bibliography, and locate the text segments where these references are cited in the text. Thus we are able to provide in the user interface counters for the number of references and in-text citations for each article, as well as pointers to the in-text citations of each reference.

We consider sentences as the basic textual unit in our processing. Our goal is to provide semantic annotations of some of the sentences and to do this we have identified a set of categories corresponding to common information needs in the context of scientific information retrieval. The semantic categories assigned to the annotated sentences can be then used to implement faceted semantic search functionalities combined with classical key-word information retrieval. Faceted search allows the user to visualize multiple categories and to filter the results according to these categories.

We segment all the paragraphs in the dataset into sentences. The segmentation process, based on the analysis of the punctuation and capitalization of the text, has already been discussed in several publications and the detailed results of the segmentation of this dataset has been given in Bertin et al. [3], using a method proposed by Mourad [7].

Our linguistic resources are based on the Contextual Exploration (CE) method described in Descles [5]. This method carries out the automatic semantic annotation of text segments for a given annotation task, such as the identification and classification of citations, the extraction of segments for summarization and the identification of specific semantic categories such as definitions, hypotheses, etc. The CE method is a decision-making procedure, presented in the form of a set of rules and linguistic clues that trigger the application of the rules. The semantic categories and the linguistic clues are organized in linguistic ontologies that correspond to the annotation tasks.

We have annotated the sentences in our corpus with a set of categories that correspond to common semantic relations expressed in scientific articles:

result: sentences that express a result obtained by the paper or by cited papers.

summarize: sentences that summarize a method, a paper, etc. typically found in the results and discussion sections.

scientific monitoring: sentences that express facts and speculations that are important for the monitoring of innovation and new results.

definition: sentences that express definitions given by the paper or by cited papers.

conclusion: sentences that express the conclusion of a paper.

controversy: sentences that express controversies, diverging opinions, etc.

agreement: sentences that express agreement in the methods, results, etc. of a paper and of cited papers.

opinion: sentences that express opinions of the authors of a paper.

The eight semantic categories are not equally represented in the corpus. Figure 1 presents the relative percentage of sentences annotated by each semantic category. The majority of annotated sentences were categorized as *result*, *summarize* and *scientific monitoring*, and these three categories account for more than 75 % of the annotations. The categories expressing opinions and subjective

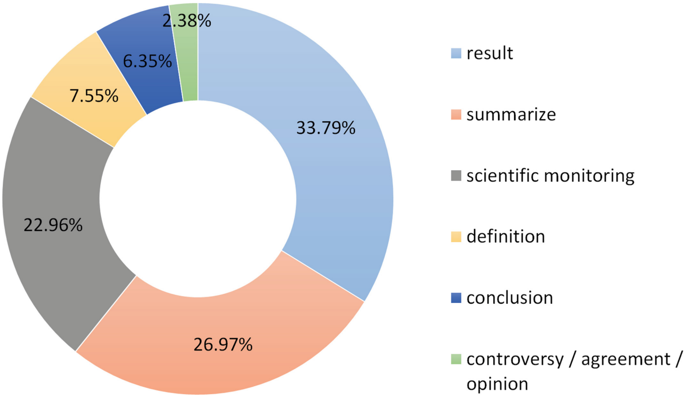


Fig. 1. Annotations by semantic category

Table 2. Semantic annotations

Journal	Articles with annotations	Annotated sentences
PLOS Biology	1,157	1,654
PLOS Computational Biology	1,440	2,782
PLOS Genetics	1,644	2,428
PLOS Medicine	635	778
PLOS Neglected Tropical Diseases	590	752
PLOS Pathogens	1,459	2,408
PLOS ONE	18,419	26,855
Total	*25,344*	*37,657*

evaluations of previous research, *controversy*, *agreement* and *opinion*, are less frequent in the corpus (about 2.4 % of the annotated sentences), as could be expected for scientific writing.

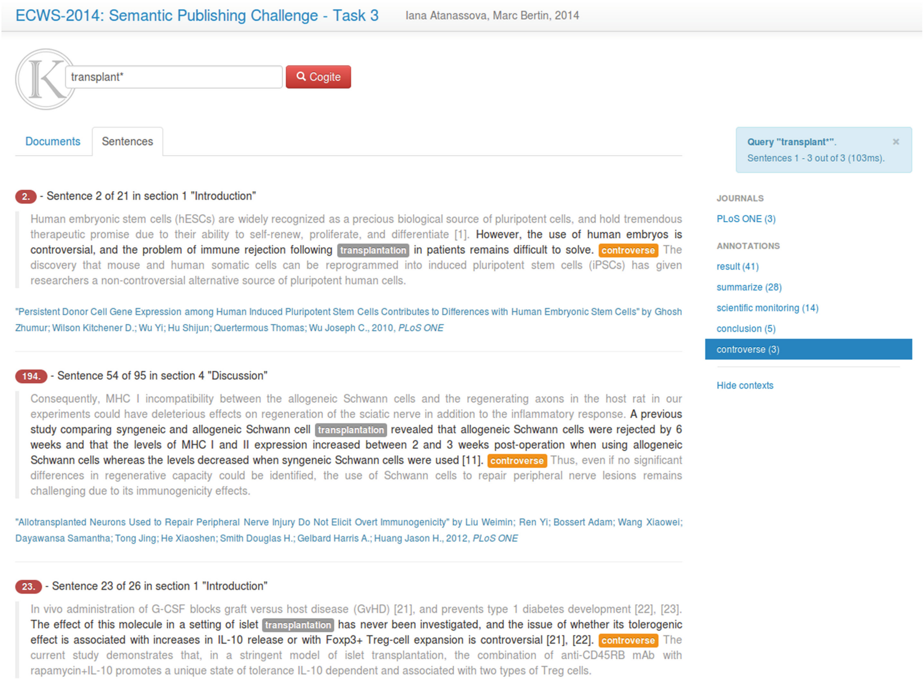
Table 2 presents the number of articles containing annotations and the number of annotated sentences. We have not evaluated the annotations for this dataset. Previous works [2] have provided evaluations of the annotation methodology working on other datasets and have obtained rather high precision values. The annotations can be converted into Linked Data using machine-readable RDF for interoperability with other tools. Our results can be used to provide an annotated corpus for the development of other approaches, for example using name-spaces and already existing vocabularies such as SPAR and DoCO [9].

3 Semantic Search Engine

We have implemented a semantic search engine using Apache Solr Search Server. The annotated XML documents were indexed using XSLT import handles. Solr uses the Lucene Java search library for full-text indexing and search. We have indexed both the articles and the sentences as two different document types that are linked in Solr's index. All annotated sentences were indexed together with their annotation categories and with their immediate context (previous and next sentence).

The search interface provides search on two levels, documents and sentences. On each level, the semantic annotations are visible and can be used as facets in order to filter the results. The initial result list is obtained by keyword search. Classical query syntax (use of *, AND, OR, etc.) is supported by Solr's query parser.

On the document level, the user has access to the list of relevant papers. Each paper is presented by its metadata. Two new types of information are given compared to classical document search: the annotations in the paper (categories



2. - Sentence 2 of 21 in section 1 "Introduction"

Human embryonic stem cells (hESCs) are widely recognized as a precious biological source of pluripotent cells, and hold tremendous therapeutic promise due to their ability to self-renew, proliferate, and differentiate [1]. However, the use of human embryos is controversial, and the problem of immune rejection following transplantation in patients remains difficult to solve. controverse The discovery that mouse and human somatic cells can be reprogrammed into induced pluripotent stem cells (iPSCs) has given researchers a non-controversial alternative source of pluripotent human cells.

"Persistent Donor Cell Gene Expression among Human Induced Pluripotent Stem Cells Contributes to Differences with Human Embryonic Stem Cells" by Ghosh Zhumur; Wilson Kitchener D.; Wu Yi; Hu Shijun; Quertermous Thomas; Wu Joseph C., 2010, *PLoS ONE*

194. - Sentence 54 of 95 in section 4 "Discussion"

Consequently, MHC I incompatibility between the allogeneic Schwann cells and the regenerating axons in the host rat in our experiments could have deleterious effects on regeneration of the sciatic nerve in addition to the inflammatory response. A previous study comparing syngeneic and allogeneic Schwann cell transplantation revealed that allogeneic Schwann cells were rejected by 6 weeks and that the levels of MHC I and II expression increased between 2 and 3 weeks post-operation when using allogeneic Schwann cells whereas the levels decreased when syngeneic Schwann cells were used [11]. controverse Thus, even if no significant differences in regenerative capacity could be identified, the use of Schwann cells to repair peripheral nerve lesions remains challenging due to its immunogenicity effects.

"Allotransplanted Neurons Used to Repair Peripheral Nerve Injury Do Not Elicit Overt Immunogenicity" by Liu Weimin; Ren Yi; Bossert Adam; Wang Xiaowei; Dayawansa Samantha; Tong Jing; He Xiaoshen; Smith Douglas H.; Gelbard Harris A.; Huang Jason H., 2012, *PLoS ONE*

23. - Sentence 23 of 26 in section 1 "Introduction"

In vivo administration of G-CSF blocks graft versus host disease (GvHD) [21], and prevents type 1 diabetes development [22], [23]. The effect of this molecule in a setting of islet transplantation has never been investigated, and the issue of whether its tolerogenic effect is associated with increases in IL-10 release or with Foxp3+ Treg-cell expansion is controversial [21], [22]. controverse The current study demonstrates that, in a stringent model of islet transplantation, the combination of anti-CD45RB mAb with rapamycin+IL-10 promotes a unique state of tolerance IL-10 dependent and associated with two types of Treg cells.

Fig. 2. Semantic search interface - sentence level search

and sentences extracted from the document) and some statistics about the article (numbers of references, number of in-text citations, etc.).

On the sentence level, as shown on Fig. 2, the search results are given as a list of annotated sentences in their contexts (previous and next sentence in the same paragraph). A sentence is considered as relevant if it contains the keywords and is annotated with one of the semantic categories that the used has selected as filters. For each sentence, the interface provides additional information for its position in the paper (the first number that appears in a red bullet), its position in the section and the bibliographic information of the paper.

The interface is available on http://sempub2014.nlp-labs.org/task3/.

4 Discussion and Conclusion

The semantic facets that we propose enable the user to filter the results according to a set of semantic categories. The annotations that generate the semantic facets are obtained using resources, such as linguistic clues and rules, and can be viewed as complex query patterns that, combined with keyword search, allow the user to access specific types of information in scientific papers. Thus, the semantic facets provide the possibility to identify highly relevant sentences among the results of keyword search. Furthermore, the automatic semantic annotation approach also allows the generation of Linked Open Data in order to propose semantic resources that can be used by different systems for the purpose of scientific knowledge extraction.

Our demonstrator presents a first implementation of an information retrieval system using semantic facets on the sentence level. This approach provides a new way to navigate in scientific papers and access relevant information. Further improvements can be made in the segmentation and annotation processing. This online version is an early prototype and our goal is to develop other semantic categories and facets related to scientific articles.

Acknowledgments. We thank Benoît Macaluso of the Observatoire des Sciences et des Technologies (OST), Montreal, Canada, for harvesting and providing the PLOS dataset.

References

1. Bertin, M., Atanassova, I.: Semantic enrichment of scientific publications and metadata : citation analysis through contextual and cognitive analysis. In: Proceedings of the 1st International Workshop on Mining Scientific Publications, in Conjunction with Joint Conference on Digital Libraries JCDL-2012. ACM/IEEE (2012)
2. Bertin, M., Atanassova, I., Desclés, J.P.: Automatic analysis of author judgment in scientific articles based on semantic annotation. In: Proceedings of the 22nd International Florida Artificial Intelligence, Research Society Conference, Sanibel Island, Florida. pp. 19–21 (2009)
3. Bertin, M., Atanassova, I., Lariviere, V., Gingras, Y.: The distribution of references in scientific papers: an analysis of the IMRaD structure. In: 14th International Society of Scientometrics and Informetrics Conference, pp. 591–603. International Society for Informetrics and Sciento (2013)
4. Buscaldi, D., Zargayouna, H.: Yasemir: yet another semantic information retrieval system. In: Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval, pp. 13–16. ACM (2013)
5. Desclés, J.P.: Contextual exploration processing for discourse and automatic annotations of texts. In: FLAIRS Conference, pp. 281–284 (2006)
6. Liakata, M., Thompson, P., de Waard, A., Nawaz, R., Maat, H.P., Ananiadou, S.: A three-way perspective on scientific discourse annotation for knowledge extraction. In: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, pp. 37–46. Association for Computational Linguistics (2012)
7. Mourad, G.: La segmentation de textes par exploration contextuelle automatique, présentation du module segatex. ISLsp, Inscription Spatiale du Langage : structure et processus IRIT, Université Paul Sabatier, Toulouse (2002)
8. Novacek, V., Groza, T., Handschuh, S., Decker, S.: Coraal - dive into publications, bathe in the knowledge. Web Semant. Sci. Serv. Agents World Wide Web **8**(2–3), 1–10 (2010)
9. Shotton, D., Peroni, S.: DoCO, the document components ontology (2011)
10. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, pp. 103–110. Association for Computational Linguistics, Stroudsburg, PA, USA (2006)
11. You, W., Fontaine, D., Barthès, J.P.: An automatic keyphrase extraction system for scientific documents. Knowl. Inf. Syst. **34**(3), 691–724 (2013)