Faceted Semantic Search for Scientific Papers

Iana Atanassova and Marc Bertin

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

Abstract. Information Retriveal in scientific publications can be enhanced through semantic annotation in view to identify relevant sentences containing specific semantic relations such as the expression of methods, definitions, hypotheses. We present an IR System for scientific publications that provides the possibility to filter results according to semantic facets. The semantic annotations are obtained using a rule-based method that identifies specific linguistic clues organized into a linguistic ontology. This approach is implemented with Solr Search Server and offers efficient search and navigation in scientific papers. The demontrator is available at http://sempub2014.nlp-labs.org/task3/.

Keywords: Semantic Annotation, Information Retrieval, Faceted Search, Semantic Facets, Solr

1 Introduction

Today, with the emergence of open science, more scientific articles are available in full text. The volume of textual data provided, fosters the development of new tools to effectively explore the content of research papers. The exploitation of semantic annotation for information retrieval is the subject of many papers (e.g. [6]) and the extraction of keyphrases from scientific articles (see [9]) is a closely related subject. The goal of the development of new information retrieval tools is to diminish the mental workload of users in the production of mental representations of documents in order to identify realevant information. This point of view is discussed by Bertin and Atanassova [2].

In this paper, we describe a semantic search engine that provides a new way to access relevant information in scientific papers. To implement the automatic annotation of the set of semantic categories in texts we use the work of [4]. Other applications of this approach are shown in [1,3].

2 Method

For this study, we have processed articles from seven scientific 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)[1], providing 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 shows the number of articles and sentences processed for each journal.

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*

Table 1. Dataset

We consider sentences as the basic unit in our processing. Our goal is to provide semantic annotation of some of the sentences in the corpus, corresponding to specific user needs in an information retrieval context. The annotated sentences can be then used to implement semantic search functionalities combined with classical key-word information retrieval. Faceted search allows the user to visualize multiple categories and to filter the reuslts according to these categories. Figure 1 presents the processing steps that will be described below. The interface is available on `http://sempub2014.nlp-labs.org/task3/`.

Metadata extraction. 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 citations of each reference.

Segmentation. We segment all the paragraphs in the dataset into sentences. The segmentation process, based on the analysis of the punctuation, has already been discussed in several publications and the detailed results of the segmentation of this dataset has been given in Bertin et al. [5].

Semantic annotation. Our linguistic resources are based on the Contextual Exploration (CE) method described in Descles [7]. This method carries out the

[1] This Standard is an application of NISO Z39.96-2012 and JATS is a continuation of the NLM Archiving and Interchange DTD: `http://jats.nlm.nih.gov`

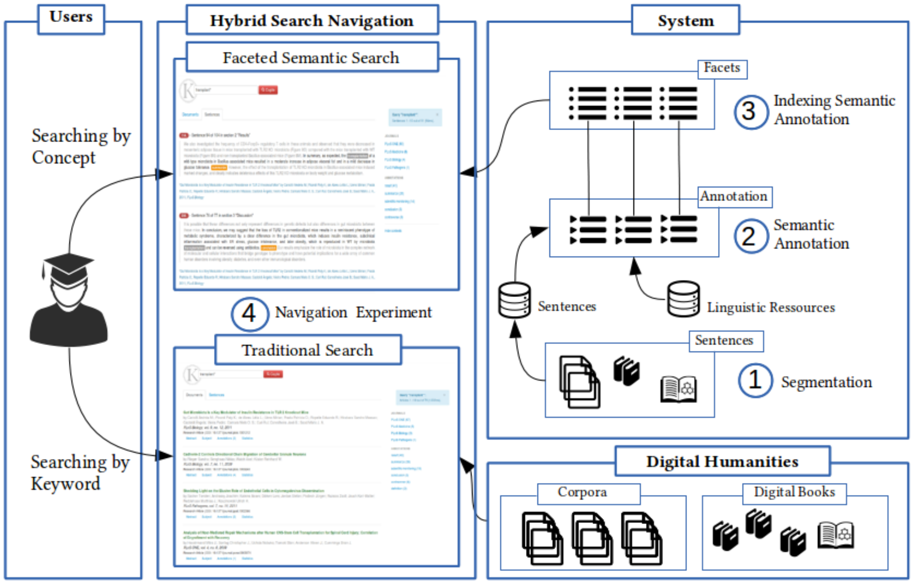

Fig. 1. Faceting & semantic annotation

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.

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.
- *secientific 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.
- *controverse*: sentences that express controverses, 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.

Table 2 presents the number of articles containing annotations and the number of annotated sentences. We have not evaluated the annotations for this

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*

Table 2. Annotation statistics

dataset. Previous works [3,4] 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. After evaluation, we will be able to provide an annotated corpus that can be used for the developement of other approaches, for example using namespaces and ontologies such as SPAR and DoCO [8].

Indexing of semantic annotations. We have implemented a semantic search engine using Apache Solr Search Server[2]. Solr provides REST-like APIs that facilitates working with XML files. In our implementation 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. This allows us to implement both article-level and sentence-level search. All annotated sentences were indexed together with their annotation categories and with their immediate context (previous and next sentence).

Semantic search. 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 acces to the list of relevant papers. Each paper is presented by its metadata. Two new types of information are given compared to classical search engines: the annotations in the paper and some statistics about the article (numbers of references and in-text citations, etc.).

On the sentence level, the search results are given as a list of sentences in their contexts, together with the annotations. A sentence is considered as relevant if it contains the keywords and is annotated with one of the semantic categories. For each sentence, we have provided some additional information for its position

[2] `https://lucene.apache.org/solr/`

in the paper (the first number that appears in a red bullet), its position in the section and the bibliographic information of the paper.

3 Conclusion and Further Development

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

Acknowledgements

We wish to thank the Observatoire des Sciences et des Technologies (OST)[3], Montreal, Canada, for harvesting and providing the PLOS dataset.

References

1. Bertin, M.: Categorizations and annotations of citation in research evaluation. In: Proceedings of the 21th International Florida Artificial Intelligence Research Society Conference. pp. 456–46 (2008)
2. 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)
3. 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)
4. Bertin, M., Atanassova, I., Desclés, J.P.: Extraction of Author's Definitions Using Indexed Reference Identification. In: Proceedings of the 1st Workshop on Definition Extraction , in conjunction with RANLP 2009. pp. 61–67. Association for Computational Linguistics (2009)
5. 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)
6. 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)
7. Desclés, J.P.: Contextual exploration processing for discourse and automatic annotations of texts. In: FLAIRS Conference. pp. 281–284 (2006)
8. Shotton, D., Peroni, S.: DoCO, the document components ontology (2011)
9. You, W., Fontaine, D., Barthès, J.P.: An automatic keyphrase extraction system for scientific documents. Knowledge and information systems 34(3), 691–724 (2013)

[3] `http://www.ost.uqam.ca/`