Knowledge Extraction and Modeling from Scientific Publications

Francesco Ronzano and Horacio Saggion

Natural Language Processing Group (TALN)
Universitat Pompeu Fabra, Barcelona, Spain
{francesco.ronzano,horacio.saggion}@upf.edu

Abstract. During the last decade the amount of scientific articles available online has substantially grown in parallel with the adoption of the Open Access publishing model. Nowadays researchers, as well as any other interested actor, are often overwhelmed by the enormous and continuously growing amount of publications to consider in order to perform any complete and careful assessment of scientific literature. As a consequence, new methodologies and automated tools to ease the extraction, semantic representation and browsing of information from papers are necessary. We propose a platform to automatically extract, enrich and characterize several structural and semantic aspects of scientific publications, representing them as RDF datasets. We analyze papers by relying on the scientific Text Mining Framework developed in the context of the European Project Dr. Inventor. We evaluate how the Framework supports two core scientific text analysis tasks: rhetorical sentence classification and extractive text summarization. To ease the exploration of the distinct facets of scientific knowledge extracted by our platform, we present a set of tailored Web visualizations. We provide on-line access to both the RDF datasets and the Web visualizations generated by mining the papers of the 2015 ACL-IJCNLP Conference.

Keywords: scientific knowledge extraction, knowledge modeling, RDF, software framework

1 Introduction: dealing with scientific publications overload

Currently, researchers have access to a huge and rapidly growing amount of scientific literature available on-line. Recent estimates reported that a new paper is published every 20 seconds [1]. PubMed[1], the reference publication index for life science and biomedical topics, currently includes about 24.6 million papers

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the European Project Dr. Inventor (FP7-ICT-2013.8.1 - Grant no: 611383).

[1] http://www.ncbi.nlm.nih.gov/pubmed

with a growth rate of about 1,370 new articles per day. Elsevier Scopus[2] and Thomson Reuters ISI Web of Knowledge[3], the two biggest privately held journal indexes, respectively contain more than 57 and 90 million papers.

At the same time, during the last few years the number of scientific papers that are freely accessible on-line considerably grew [2, 3]. Currently, the Directory of Open Access Journals[4], one of the most authoritative indexes of high quality, Open Access, peer-reviewed publications, lists more than 10,800 journals and 2.1 million papers. The full text of 27% of the articles indexed by PubMed is available on-line for free. Sometimes between 2017 and 2021, more than half of the global papers are expected to be published as Open Access articles [4].

The exploration of recent advances concerning specific topics, methods and techniques, peer reviewing, the writing and evaluation of research proposals and in general any activity that requires a careful and comprehensive assessment of scientific literature has turned into an extremely complex, time-consuming task.

In this context, considering also the increasing amount of scientific information freely accessible on-line, the availability of text mining tools able to extract, aggregate and turn scientific unstructured textual contents into well organized and interconnected knowledge is fundamental. However, scientific publications are characterized by several structural, linguistic and semantic peculiarities: general purpose text mining tools and techniques often need to be substantially adapted and extended in order to correctly deal with their contents. Even if the adoption of Web-friendly, textual formats and XML dialects like JATS[5] [5], Elsevier Schemas[6] and RASH[7] is rapidly spreading, *the majority of scientific papers is still available as PDF documents*, thus requiring proper tools to consistently extract their contents [6, 8, 9]. Scientific publications include *common structural elements* (title, authors, abstract, sections, figures, tables, citations, bibliography) that often requires customized approaches to be properly characterized [10–13]. Similarly, scientific articles are also distinguished by their *peculiar discursive structure* (background, challenge, outcome, future works) [14, 15]. Papers are interconnected by their *network of citations* that constitutes the basis of widespread count-based metrics (i.e. h.index). Citation semantics has started to be exploited in several contexts including opinion mining [16, 17] and scientific text summarization [18, 19]. Integrated scientific article mining systems have been proposed and released in order to perform complex paper analysis tasks like the joint annotation of several kinds of structural information [25] or the semantic characterization and querying of contents [22].

Recently, in parallel to the diffusion of new approaches to scientific text mining, several investigation and development efforts have also been focused on the modeling and interlinking of scholarly publishing contents by relying on Semantic

[2] http://www.scopus.com/
[3] http://www.webofknowledge.com/
[4] https://doaj.org/
[5] http://jats.nlm.nih.gov/
[6] http://www.elsevier.com/author-schemas/elsevier-xml-dtds-and-transport-schemas
[7] https://rawgit.com/essepuntato/rash/master/documentation/index.html

Web standards and technologies [20–22]. This trend is usually referred to as semantic publishing [23]. In this context, the Semantic Publishing Challenges [24], organized as part of the Extended Semantic Web Conferences, represents an important discussion and evaluation venue.

In this paper, we present a platform that extracts semantically rich information from scientific articles and represents it both as RDF datasets and by means of properly tailored Web visualizations. To mine the contents of scientific publications, we rely on the Text Mining Framework developed in the context of the European Project Dr. Inventor. The Framework integrates several text mining modules that spot many structural and semantic facets of scientific publications. In comparison with existing tools, the Dr. Inventor Text Mining Framework provides a coherent system that enables the automated extraction of a greater and richer set of structural and semantic knowledge facets from scientific articles. Besides the identification and enrichment of papers' structural and citation-related data, by relying on the Framework it is possible to perform the automated rhetorical classification of sentences, the disambiguation and entity linking of papers' contents and the creation extractive summaries of an article. Moreover it enables the creation of subject-verb-object graph representations of an article that are being exploited in the context of the Dr. Inventor Project to identify creative analogies across papers [39]. The Framework is distributed as a self-contained Java library[8], thus providing a convenient tool both to bootstrap more complex scientific publication analysis experiments as well as to foster the creation of structured, semantically-rich knowledge from papers' contents.

In Section 2 we describe the main scientific text mining modules that compose the Framework. Section 3 provides an evaluation of the performances of the Framework with respect to the identification of the discourse rhetorical category of sentences and the selection of the most relevant sentences to summarize a paper. In Section 4 we outline our approach to the representation of the contents mined from a paper as an RDF dataset. Section 5 introduces a set of Web visualizations useful to provide an easy and interactive way to explore the information extracted from a scientific publication. In Section 6 we present our conclusions and sketch future venues of research.

2 Exploiting the Dr. Inventor Framework to mine scientific publications

We rely on the Dr. Inventor Text Mining Framework [26] (DRI Framework) to extract and enrich the information necessary to generate both RDF datasets and Web visualizations from scientific publications. The DRI Framework integrates, extends and customizes a collection of scientific text mining modules and services in order to support the joint analysis of structural, linguistic and semantic aspects of scientific publications. It has been implemented and is distributed as a Java library. The DRI Framework relies on the GATE Text Engineering

[8] http://backingdata.org/dri/library/

Platform [27] as a common glue to integrate its text mining modules. Figure 1 provides an overview of the modules integrated in the current version of the DRI Framework. Each one of them is described in greater detail in the remaining part of this Section.

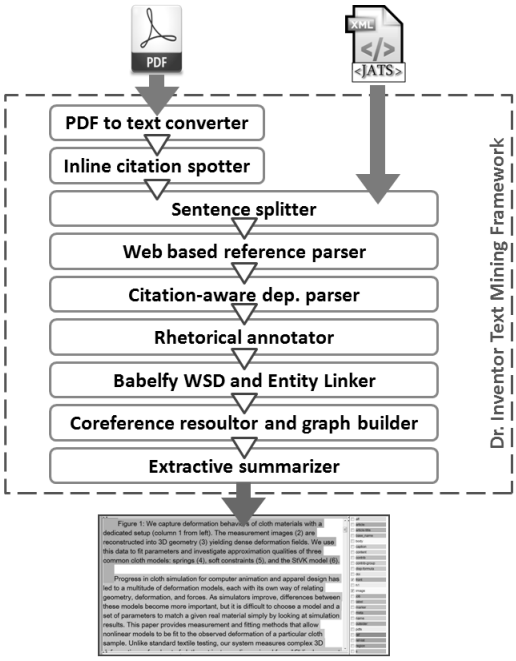

Fig. 1. Architectural overview of the modules of the Dr. Inventor Framework.

The DRI Framework can mine scientific publications both in PDF and JATS XML format. As shown in Figure 1, two additional text mining modules are needed to process the contents of PDF articles, with respect to publications available as JATS XML files. The first module is the **PDF to text converter** that extracts textual contents from PDF documents. After a comparative analysis and evaluation of several PDF-to-text conversion approaches both generic and customized to scientific publications, we decided to rely on PDFX and its Web API[9] [6] to convert PDF files to text. PDFX is a rule-based PDF mining engine that enables most of the times the extraction of clean and consistent semi-structured textual contents form the PDF file of a scientific article. We rely on the structured XML output of PDFX to identify the title, the abstract, the sections and the bibliographic entries of a paper.

Once PDF papers are converted to text, the **Inline citation spotter** module is executed. By means of a set of JAPE rules [37] covering several widespread ci-

[9] http://pdfx.cs.man.ac.uk/

tation styles, inline citation spans and inline citation markers are identified inside the textual contents of a paper (Figure 2-a). Then each inline citation marker is linked to the related bibliographic entry (bibEntry) by a set of heuristics tailored to the detected inline citation style (Figure 2-b).

The **PDF to text converter** module and the **Inline citation spotter** module are not needed for publications available as JATS XML files since their XML markup already identifies the structural elements just contemplated (sections, citations and related bibEntries).

The **Sentence splitter module** identifies the sentence boundaries inside each article by relying on a rule-based sentence splitting approach [27] that has been customized so as to deal with some peculiarity of scientific publications (expressions like: i.e., et. al., Fig., Tab. that do not identify the end of a sentence).

The **Web based reference parser** analyzes the contents of each bibEntry in order to identify its structural components (like paper title, authors, publication year, etc.). It also retrieves references to those bibEntries from external publication indexes (Figure 2-c) by querying and merging the results of the online Web APIs exposed by *Bibsonomy*[10], *CrossRef*[11] and *FreeCite*[12].

At this stage, every sentence of the paper is processed by means of the next two modules. First of all the **Citation-aware dependency parser** performs the tokenization, lemmatization and POS-tagging of each sentence and builds a dependency tree by relying on a modified version of the MATE tools[13] [7] that has been properly customized to correctly deal with inline citation spans. When an inline citation span has a syntactic role inside the sentence where it occurs, it is considered as a single word when building the dependency tree of the sentence (Figure 2-d, first example). On the contrary, when the inline citation span has not syntactic role in the sentence, it is ignored (Figure 2-d, second example). The upper part of 2-e shows the POS tags and the dependency tree of a sentence in which the subject is the inline citation span *(Hu, 2004)*.

Thanks to the sentence analysis performed by the Citation-aware dependency parser, the **Rhetorical annotator** processes the contents of each sentence to identify its scientific discourse rhetorical category (see [29] for details on the annotation schema) among: Approach, Challenge, Background, Outcomes and Future Work. This module relies on a Logistic Regression classifier trained on the manual annotations of Dr. Inventor Corpus[14] [29, 30]. In Section 3 we provide more details on the Corpus and evaluate the performance of this module.

The next module of the DRI Framework is the **Babelfy WSD and Entity Linker**. It processes the contents of the paper by invoking the Babelfy Web API[15] [33]. Babelfy is a graph-based methodology to perform Entity Linking

[10] http://www.bibsonomy.org/help/doc/api.html
[11] http://search.crossref.org/help/api
[12] http://freecite.library.brown.edu/welcome
[13] https://code.google.com/p/mate-tools/
[14] http://sempub.taln.upf.edu/dricorpus
[15] http://babelfy.org/

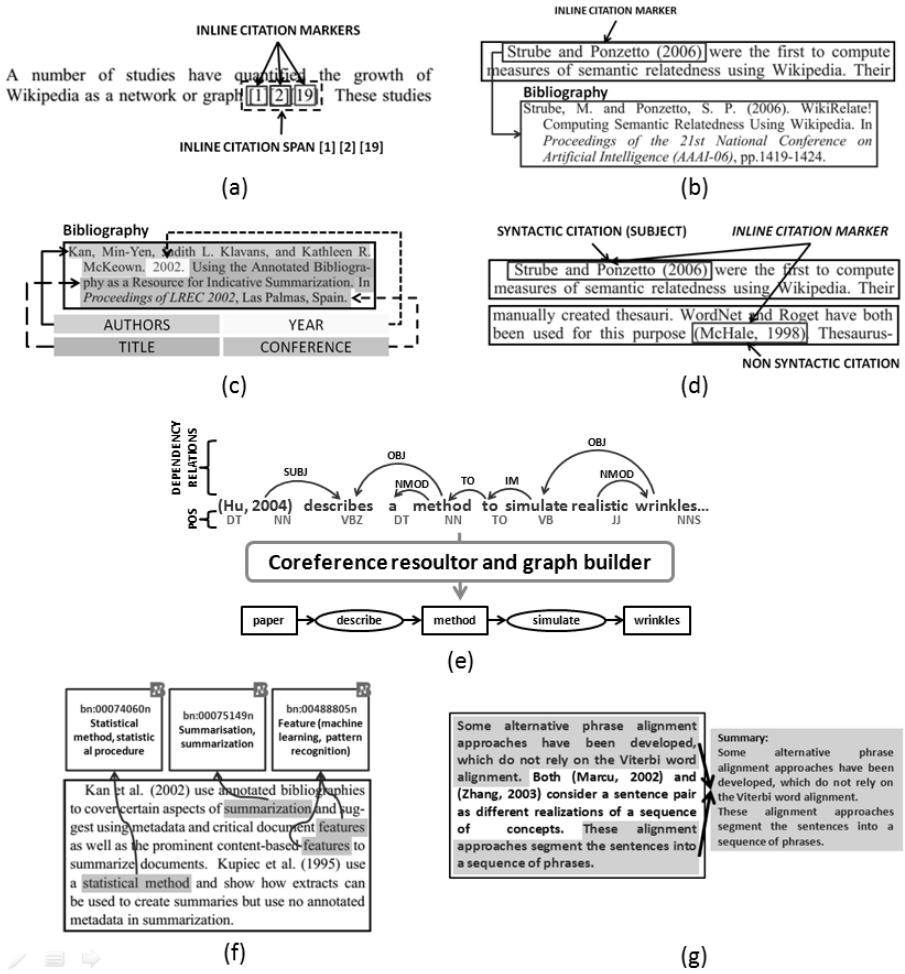

Fig. 2. Functional schemas of the modules of the Dr. Inventor Framework.

and Word Sense Disambiguation, relying on the Babelnet semantic network[16]. Thanks to Babelfy the occurrences of concepts and Named Entities are spotted inside the text of each paper and properly linked to their right meaning chosen in the sense inventory of Babelnet. Figure 2-f shows a portion of an article where the occurrences of three concepts (*summarizaiton*, *features* and *statistical method*) have been spotted and linked to their respective Babelnet synsets (senses).

The **Coreference resolutor and graph builder** module, starting from the outputs of the Citation-aware dependency parser, represents each sentence of a paper as a Subject-Verb-Object graph. An example of such graph is shown in

[16] http://babelnet.org/

Figure 2-e. A rule-based nominal and pronominal coreference resolutor has been implemented in this module in order to support the integration of Subject-Verb-Object graphs generated from distinct sentences by merging the nodes that refer to the same entity. For instance, the coreference resolutor is able to spot that a pronominal node refers to a specific nominal entity, thus merging of both nodes.

The last module of the DRI Framework is the **Extractive summarizer**. It implements extractive paper summarization algorithms thanks to the integration of the SUMMA toolkit [34][17]. These algorithms rate the sentences of a paper with respect to their relevance for the inclusion in a summary: the top-n rated sentences are then chosen and composed so as to generate the extractive summary of the article (Figure 2-g). The current version of the DRI Framework implements two basic sentence ranking approaches: the sentence similarity with the title of the paper and the sentence similarity with the centroid of each section of the paper. In Section 3 we evaluate the performance of distinct summarization approaches including the ones implemented by this module.

The DRI Framework is distributed as a self-contained Java library that exposes a convenient API in order to invoke the execution of the scientific text mining modules described in this Section. The results of the paper analyses can be easily accessed thanks to the tailored object-oriented data model of scientific publication that is implemented by the DRI Framework. The last version of the DRI Framework as well as the related JavaDoc, tutorials and code examples can be accessed online at: http://backingdata.org/dri/library/.

3 Evaluation of rhetorical sentence annotation and extractive summarization

In this Section we present two experiments useful to measure the performance of two core modules of the DRI Framework: the Rhetorical annotator and the Extractive summarizer. Both experiments rely on the textual annotations of the Dr. Inventor Corpus. This Corpus includes 40 Computer Graphics papers containing 8,877 sentences that have been manually annotated with respect to their scientific discourse rhetorical category. Moreover, the corpus includes for each paper three handwritten summaries of maximum 250 words.

The **Rhetorical annotator** module integrated in the DRI Framework is based on a Logistic Regression rhetorical sentence classifier implemented by relying on the Weka data mining tools [38]. To select the best approach to determine the rhetorical category of each sentence, we compared the performance of two classifiers: Support Vector Machine (SVM) with linear kernel [28] and Logistic Regression. We represent each sentence to classify by means of a set of lexical and semantic features and evaluate each classification approach by performing a 10-fold cross validation over the 8,877 manually annotated sentences of Dr. Inventor Corpus [29]. The results are shown in Table 1 where we can notice that the Logistic Regression performs better than the SVM classifier both on average

[17] http://www.taln.upf.edu/pages/summa.upf/

and with respect to each rhetorical category. In general, the performance of the classifier for each rhetorical category decreases with respect to the frequency of annotated sentences belonging to that category in the Dr. Inventor Corpus.

Rhetorical Category	**Logistic Regression**	**SVM**
Approach	0.876	0.851
Background	0.778	0.735
Challenge	0.466	0.430
Future Work	0.675	0.496
Outcome	0.679	0.623
Avg. F1:	**0.801**	**0.764**

Table 1. F1 score of Logistic Regression and SVM classifier evaluated by a 10-fold cross validation over the manually annotated sentences of Dr. Inventor Corpus.

The **Extractive summarizer** module implements distinct approaches to rank the sentences of a paper with respect to their relevance to be included in a summary. In the rest of this Section we compare the summarization performances three approaches: the two summarization techniques implemented by the DRI Framework (sentence similarity with the title of the paper and sentence similarity with the centroid of each section of the paper) and the TextRank graph-based summarization algorithm [31].

To this purpose, for each paper of Dr. Inventor Corpus we generate three summaries of approximately 250 words, each one by relying on a specific summarization approach. Then, we compare each automatically generated summary with the three human handwritten ones by computing the average ROUGE-2 score[18] [32]. For each summarization approach, we determine the global ROUGE-2 score by computing the average ROUGE-2 of all the 40 papers of Dr. Inventor Corpus. In this way we can quantify and compare the performance of each summarization approach. By scoring sentences with respect to their similarity with the title, we obtain a global ROUGE-2 score of 0.3151 that improves up to 0.3427 when we score sentences by considering their similarity with each section centroid. The best summarization performance (global ROUGE-2 0.3617) is obtained by relying on the TextRank algorithm. We are planning to integrate this algorithm in the next releases of the DRI Framework.

4 Semantic modeling of scientific publications

In this Section we present our approach to model as RDF data the structural and semantic information mined by means of the DRI Framework, thus enabling

[18] Rouge-2 is a measure which compares n-grams in automatic summaries to n-grams in gold stadard summaries

the automated creation of structured, rich data collections describing scholarly contents, in accordance to the semantic publishing principles. We extend and enrich the basic RDF data modeling approach of scientific papers we adopted in the context of our participation to the Semantic Publishing Challenge 2015 [35]. In particular, our RDF data modeling choices have been driven by the necessity to represent the varied information that can be mined from a publication thanks to the DRI Framework. Thus, besides the representation of articles' metadata and bibliographic entries, our RDF data model contemplates the possibility to describe the structure of a paper, by identifying its abstract, sections and sentences. Each sentence can be characterized by both its rhetorical category and the Babelnet synsets (senses) that have been spotted inside its content. Moreover we link each bibEntry to all the sentences that include the related in-line citations.

The DRI Framework Java library has been properly extended with methods useful to trigger the automatic generation of the RDF dataset of a paper. The RDF datasets generated from the papers presented at the 2015 ACL-IJCNLP Conference can be downloaded online[19]. In the remaining part of this Section we describe in more detail the RDF data modeling choices we made and the ontologies we reused and extended to represent the contents of scientific publications.

Figure 3 schematizes our RDF model of scientific articles. We relied on the core RDF data modeling approaches, patterns and ontologies accessible in the Semantic Publishing and Referencing (SPAR) Portal[20] [36]. The SPAR Portal defines and documents a complete and consistent set of 12 ontologies tailored to model several aspect of scientific publishing, including articles' metadata, authors, bibliography, citations, publication workflows, etc. From the classes and the properties modeled by the SPAR ontologies, we reused and derived - in the dri namespace - new sub-classes and sub-properties. As a consequence, we included the related T-BOX axioms in the RDF Datasets we generate. The URIs needed to unambiguously reference each article together with its components (authors, sections, sentences, bibEntries, etc.) are instantiated in a namespace provided by DRI Framework users.

Figure 3-a shows how we represent the structured contents of a paper as RDF triples. Two URIs are generated to reference the abstract and the body of the paper (respectively the *FrontMatter_URI* and the *BodyMatter_URI* in Figure 3-a). Both the abstract and the body may contain a list of sections (*IntroSection_URI* and *MethodSection_URI* in Figure 3-a). Each section is identified by an URI and related to an instance of the doco:SectionTitle class that represents its title. The abstract, body or sections of the paper can contain one or more sentences, each one identified by an URI (*Sentence1_URI*, ..., *SentenceN_URI* in Figure 3-a). The lower part of Figure 3-a shows the association of the sentences of the paper to their scientific discourse rhetorical category. This is achieved by representing the corresponding *sentence_URI* as an instance of one of the following classes: dri:Approach, dri:Challenge, dri:Background, dri:Outcomes and

[19] Download link: http://backingdata.org/dri/viz/
[20] http://www.sparontologies.net/

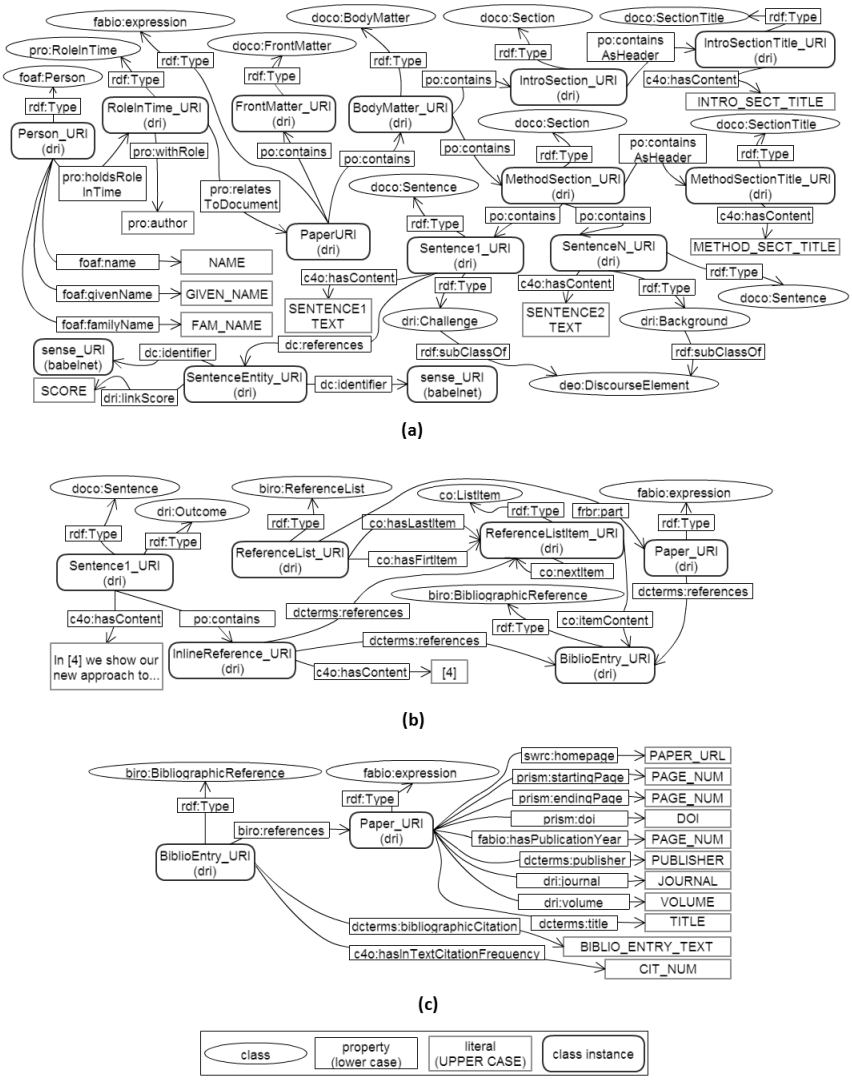

Fig. 3. RDF data model of scientific article: a) authors and internal structure of the paper including sections and sentences with their rhetorical class and associated Babelnet senses; b) list of bibliographic entries of the paper together with the pointer to the sentences in which each bibliographic entry occurs; c) descriptive data of both papers and bibliographic entries. Ontology prefixes: **doco** *Document Components Ontology*, **fabio** *FRBR-aligned Bibliographic Ontology*, **c4o** *Citation Counting and Context Characterization Ontology*, **pro** *Publishing Roles Ontology*, **biro** *Bibliographic Reference Ontology*, **swrc** for the Semantic Web for Research Communities Ontology, **prism** *PRISM Metadata Ontology*, **foaf** *Friend Of A Friend Ontology*, **po** *Pattern Ontology*, **co** *Collections Onology*, **dc** and **dcterms** *Dublin Core Ontology*. The prefix **dri** identifies the classes and properties of Dr. Inventor Ontology.

dri:FutureWork. The association of a Babelnet synset (sense) to the sentence where the same synset has been spotted is modeled by linking the URI of the sentence to a *SentenceEntity_URI*. The *SentenceEntity_URI* is in turn characterized by the URIs of both the Babelnet synset and the DBpedia entity that represent that sense. Moreover, each association of a sense to a sentence is characterized by a score (literal object of the property dri:linkScore). This score is a double value that provides an estimate of the strength of the concept-to-sentence association.

On the left side of Figure 3-a, we show how the Publishing Roles Ontology is exploited in order to model the authors of a paper. The same ontology is also used to represent the editors of an article.

Figure 3-b and Figure 3-c show the RDF representation of the bibliography of a paper. By relying on the Collections Ontology, the bibEntries are represented as an ordered list. An URI is assigned to each inline citation belonging to a specific sentence of the paper (*InlineRefrence_URI* in Figure 3-b). The inline reference URI relates the sentence that contains the inline citation to the referenced bibEntry. Also the textual contents of the inline reference are specified by means of the property c4o:hasContent. Figure 3-c shows how each bibEntry is characterized by specifying the cited paper (identified by its URI, *Paper_URI*), the text of the same bibEntry and the number of times that bibEntry is cited inside the considered paper.

When we generate these data our focus is put on the creation of a consistent and semantically-rich RDF representation of the contents mined from a single scientific publication by means of the DRI Framework. As far as concern the creation of links to external Linked Data, the RDF datasets we generate connect publications and bibEntries to bibliographic indexes like Bibsonomy and relates each sentence of the paper to the Babelnet synsets (senses) mentioned in its contents. We plan to extend our RDF generation approach so as to foster the creation of new, richer internal and external links, thus increasing data integration and interlinking.

5 Visualizing semantically enriched scientific publications

In this Section we present a set of Web visualizations we developed to support an easier and more interactive navigation of the contents mined from a scientific publication by means of the DRI Framework. The visualizations of the papers presented at the 2015 ACL-IJCNLP Conference can be accessed online[21].

The information mined from a scientific article is presented by means of a multi-tab view (see Figure 4). Each tab is meant to show a specific type of data extracted from a scientific publication together with aggregated statistical information. In the first tab, named 'Main tab' and shown in Figure 4-a, the textual content of the paper can be browsed by section. Inline citations inside each sentence of the paper can be inspected (by a click) so as to explore the detailed

[21] http://backingdata.org/dri/viz/

metadata associated by the **Web based reference parser** module. Moreover, the sentences of the paper can be highlighted in different colors with respect to the scientific discourse rhetorical category associated by the **Rhetorical annotator** module. Similarly it is also possible to highlight the sentences chosen by the **Extractive summarizer** to be part of a summary of the paper. All these features of the 'Main view' tab can be accessed by the four drop down menus that are present in its left side (Figure 4-a).

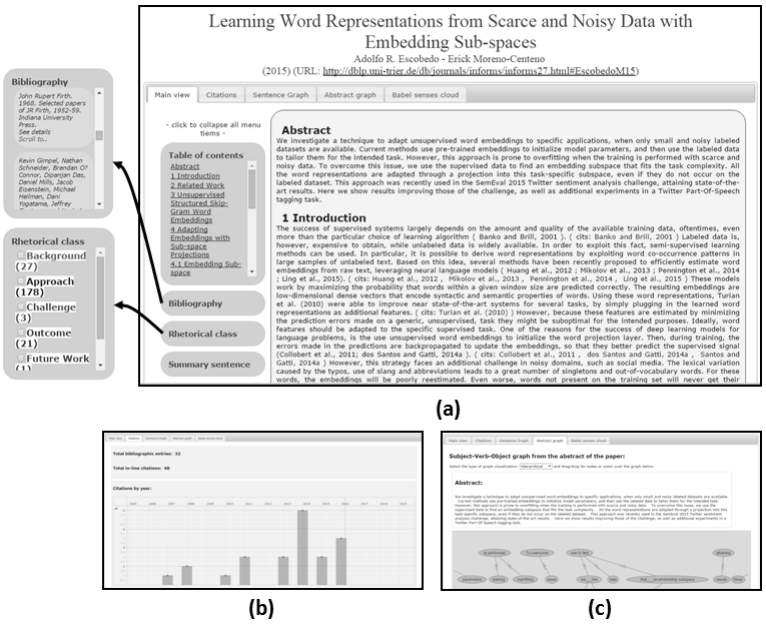

Fig. 4. Web visualizations of the information mined by the DRI Framework: (a) Main tab; (b) Citation tab; (c) Abstract graph tab.

The second tab, named 'Citations' (Figure 4-b), enables the visualization of several aggregated statistical data concerning the citations of the paper. The third and the fourth tabs enable the visualization of the Subject-Verb-Object graphs (see Figure 2-e) that represent respectively the contents of each sentence of the paper ('Sentence graph' tab) and the aggregated contents of the abstract of the paper ('Abstract graph' tab, Figure 4-c). The Subject-Verb-Object graphs are mined by both the **Citation-aware dependency parser** module and the **Coreference resoultor and graph builder** module. A fifth tab named 'Babel senses cloud' enables users to inspect the top-10 Babelnet synsets (senses) that occur in the contents of the paper, identified thanks to the **Babelfy WSD and Entity Linker** module.

6 Conclusions and future work

The amount of scientific publications available on-line is growing at an unprecedented rate together with the diffusion of the Open Access publishing model, thus turning any careful and comprehensive assessment of scientific literature into an extremely complex and time-consuming task. In this scenario, in order to help researcher and other interested actors to easily select, access and aggregate the contents of scientific papers, the availability of new approaches and tools that enable the automated analysis and interconnection of structural and semantic information from scientific literature is fundamental.

In this paper we presented a platform useful to extract several types of information from scientific publications and represent it both as RDF datasets and by means of interactive Web visualizations. In order to process, analyze and enrich the contents of a scientific article we exploited the scientific Text Mining Framework we developed in the context of the European Project Dr. Inventor. We described in detail both the scientific text analysis modules integrated into the Framework and the RDF data modeling approach we adopted. We evaluated how the framework supports rhetorical sentence classification and extractive summarization. Moreover, we presented a set of Web visualizations of the structured contents we extract from scientific articles. The Dr. Inventor Text Mining Framework is available as a self-contained Java library that provides a comprehensive, ready-to-use platform for scientific text analysis. The Framework is intended to provide an integrated tool to ease the expensive and time consuming bootstrapping of scientific text mining experiments by automatically enriching the contents of scientific papers by identifying several structural and semantic information. The Framework is also meant to foster the automated creation of scholarly publishing RDF data since it allows the creation of RDF datasets that model the knowledge mined from a paper.

As future work, we plan to further improve and extrinsically evaluate the main text analysis modules of the Text Mining Framework. In particular we plan to refine and carry out user and task-based evaluations of the Subject-Verb-Object graphs extracted from the textual contents of each paper. We are also planning to experiment new approaches to rhetorical sentence classification by relying on active learning. We would like to evaluate new ways to further characterize and take advantage of the citations of a paper by determining their polarity and purpose.

References

[1] The Rise of Open Access. Science, Vol. 342 no. 6154 pp. 58-59 - https://www.sciencemag.org/content/342/6154/58.full (2013)

[2] Bjrk, B. C., Laakso, M., Welling, P., & Paetau, P.: Anatomy of green open access. Journal of the Association for Information Science and Technology, 65(2), 237-250 (2014)

[3] Solomon, D. J., Laakso, M., & Bjrk, B. C.: A longitudinal comparison of citation rates and growth among open access journals. Journal of Informetrics, 7(3), 642-650 (2013)

[4] Lewis, D. W.: The inevitability of open access. College & Research Libraries, 73(5), 493-506 (2012)
[5] Huh, S.: Coding practice of the Journal Article Tag Suite extensible markup language. Science Editing, 1(2), 105-112 (2014)
[6] Constantin, A., Pettifer, S., & Voronkov, A.: PDFX: fully-automated PDF-to-XML conversion of scientific literature. In Proceedings of the 2013 ACM symposium on Document engineering (pp. 177-180). ACM (2013)
[7] Bohnet, B.: Very high accuracy and fast dependency parsing is not a contradiction. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 89-97). Association for Computational Linguistics (2010)
[8] Tkaczyk, D., Szostek, P., Dendek, P. J., Fedoryszak, M., & Bolikowski, L.: CERMINE–Automatic Extraction of Metadata and References from Scientific Literature. In Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on (pp. 217-221). IEEE (2014)
[9] Ramakrishnan, C., Patnia, A., Hovy, E. H., & Burns, G. A.: Layout-aware text extraction from full-text PDF of scientific articles. Source code for biology and medicine, 7(1), 7 (2012)
[10] Peng, F., & McCallum, A.: Information extraction from research papers using conditional random fields. Information processing & management, 42(4), 963-979 (2006)
[11] Do, H. H. N., Chandrasekaran, M. K., Cho, P. S., & Kan, M. Y.: Extracting and matching authors and affiliations in scholarly documents. In Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries (pp. 219-228). ACM (2013)
[12] Councill, I. G., Giles, C. L., & Kan, M. Y.: ParsCit: an Open-source CRF Reference String Parsing Package. In LREC (2008)
[13] Luong, M. T., Nguyen, T. D., & Kan, M. Y.: Logical structure recovery in scholarly articles with rich document features. Multimedia Storage and Retrieval Innovations for Digital Library Systems, 270 (2012)
[14] 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)
[15] Teufel, S.: The Structure of Scientific Articles: Applications to Citation Indexing and Summarization. Computational Linguistics, 38(2), 443-445 (2012)
[16] Nakov, P. I., Schwartz, A. S., & Hearst, M.: Citances: Citation sentences for semantic analysis of bioscience text. In Proceedings of the SIGIR04 workshop on Search and Discovery in Bioinformatics (pp. 81-88). (2004)
[17] Abu-Jbara, A., Ezra, J.,& Radev, D. R.: Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In HLT-NAACL, pp. 596-606 (2013)
[18] Abu-Jbara, A., & Radev, D.: Coherent citation-based summarization of scientific papers. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 500-509). Association for Computational Linguistics. (2011)
[19] Ronzano, F. & Saggion, H.: Taking advantage of citances: citation scope identification and citation-based summarization. Text Analytics Conference (2014)
[20] Smit, E., & Van Der Graaf, M.: Journal article mining: the scholarly publishers' perspective. Learned Publishing, 25(1), 35-46 (2012)
[21] Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F.: Semantic annotation of scholarly documents and citations. In AI* IA 2013: Advances in Artificial Intelligence (pp. 336-347). Springer International Publishing. (2013)

[22] Sateli, B., & Witte, R.: What's in this paper?: Combining Rhetorical Entities with Linked Open Data for Semantic Literature Querying. In Proceedings of the 24th International Conference on World Wide Web Companion, pp. 1023-1028 (2015)
[23] Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing, 22(2), 85-94 (2009)
[24] Di Iorio, A., Lange, C., Dimou, A., & Vahdati, S.: Semantic Publishing ChallengeAssessing the Quality of Scientific Output by Information Extraction and Interlinking. In Semantic Web Evaluation Challenges (pp. 65-80). Springer International Publishing (2015)
[25] Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, .: CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR), 18(4), 317-335 (2015)
[26] Ronzano, F., & Saggion, H.: Dr. Inventor Framework: Extracting Structured Information from Scientific Publications. Discovery Science (pp. 209-220). Springer International Publishing. (2015)
[27] Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K.: Getting more out of biomedical documents with GATEs full lifecycle open source text analytics. PLoS Comput Biol, 9(2), e1002854 (2013)
[28] Schlkopf, B., & Smola, A. J.: Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press (2002)
[29] Fisas, B., Ronzano, F., & Saggion, H.: On the Discoursive Structure of Computer Graphics Research Papers. In The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015, p. 42 (2015)
[30] Fisas, B., Ronzano, F., & Saggion, H.: A Multi-Layered Annotated Corpus of Scientific Papers. In The Language Resource and Evaluation Conference (2016)
[31] Mihalcea, R.: Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (p. 20). Association for Computational Linguistics. (2004)
[32] Lin, C. Y.: Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, Vol. 8 (2004)
[33] Moro, A., Cecconi, F., & Navigli, R.: Multilingual word sense disambiguation and entity linking for everybody. Proc. of ISWC (P&D), 25-28 (2014)
[34] Saggion, H.: SUMMA: A robust and adaptable summarization tool. Traitement Automatique des Langues, 49(2) (2008)
[35] Ronzano, F., Fisas, B., del Bosque, G. C., & Saggion, H.: On the automated generation of scholarly publishing linked datasets: the case of CEUR-WS proceedings. In Semantic Web Evaluation Challenges (pp. 177-188). Springer International Publishing. (2015)
[36] Peroni, S.: The Semantic Publishing and Referencing Ontologies. In Semantic Web Technologies and Legal Scholarly Publishing (pp. 121-193). Springer International Publishing. (2014)
[37] Thakker, D., Osman, T., & Lakin, P.: Gate jape grammar tutorial. Nottingham Trent University, UK, Phil Lakin, UK, Version, 1. (2009)
[38] Witten, I. H., & Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. (2005)
[39] O'Donoghue, D. P., Abgaz, Y., Hurley, D., Ronzano, F. & Saggion, H.: Stimulating and simulating creativity with Dr inventor. In the Proceedings of the International Conference on Computational Creativity (2015)