# MultiScien: a Bi-Lingual Natural Language Processing System for Mining and Enrichment of Scientific Collections

Horacio Saggion, Francesco Ronzano, Pablo Accuosto, and Daniel Ferrés

Large-Scale Text Understanding Systems Lab
TALN Research Group
Department of Information and Communication Technologies
Universitat Pompeu Fabra
C/Tanger 12, 08018 Barcelona, Spain
{name.surname}@upf.edu

**Abstract.** In the current online *Open Science* context, scientific datasets and tools for deep text analysis, visualization and exploitation play a major role. We present a system for deep analysis and annotation of scientific text collections. We also introduce the first version of the SEPLN Anthology, a bi-lingual (Spanish and English) fully annotated text resource in the field of natural language processing that we created with our system. Moreover, a faceted-search and visualization system to explore the created resource is introduced. All resources created for this paper will be available to the research community.

**Keywords:** Language Resources; Scientific Text Corpora; Information Extraction; Data Visualization; Semantic Analysis; PDF Conversion

## 1 Introduction

Scientific articles are among the most valuable textual records of human knowledge. In the past few years the volume of scientific publications made available online has grown exponentially, opening interesting research avenues for Natural Language Processing (NLP), Information Retrieval, and Data Visualization.

In the current online *Open Science* context [7], the availability of scientific datasets and tools for deep text analysis, visualization and exploitation plays a major role. Access to recent and past scientific discoveries, methods, and techniques is essential for scientific inquiry activities which include, among others: (i) finding open or solved problems, (ii) understanding research fields' dynamics or evolution, (iii) discovering experts in specific scientific areas, or (iv) understanding the advantages and limitations of current solutions.

In recent years, a number of initiatives have emerged to make scientific articles accessible as structured text corpora: notable examples include the ACL Anthology network [23], the Darmstadt Scientific Text Corpus [11] or the TALN Archives [4]. At the same time, efforts to make publications available through

scientific portals have proliferated with CiteSeer [15], Google Scholar [12], Microsoft Academics [29], or DBLP [14] as some of the best known. However, and to the best of our knowledge, most of these initiatives do not expose rich linguistic annotations and only perform limited analyses of the content of the scientific document, which mainly deal with the extraction of the document's logical structure and layout and the indexing of contents. Papers available as PDF documents are indeed often processed to extract relevant meta-data such as authors, titles and abstracts, addressing in some cases author disambiguation and the extraction of citations and citations' sentences for citations network building.

In order to exploit the rich content of scientific collections, we have developed a system for deep analysis of research papers. The system integrates text parsing, coreference resolution, bibliography enrichment, word sense disambiguation and entity linking, rhetorical sentence classification, citation purpose identification, open information extraction, and text summarization. Moreover, the tool is able to deal with texts in English and Spanish. For this research, we have applied our system to create the initial version of the Anthology of the Spanish Society for Natural Language Processing[1]–the SEPLN Anthology: a bi-lingual (Spanish / English) text resource in the field of Natural Language Processing built by extracting structured textual contents and performing linguistic and semantic analyses of the research articles published by SEPLN over the last years. The annotated textual contents of the SEPLN Anthology are freely available to the NLP communities in order to boost research and experimentation in the field of scientific text mining.

In order to support intelligent access to the collection, the documents, meta-data and linguistic information of this resource have been properly indexed and made available in a web platform to offer the possibility of interactively explore the knowledge from the SEPLN universe through faceted searches.

We summarize the contributions of this paper as follows:

- The first bi-lingual scientific text analysis system which is specially adapted to SEPLN and other types of scientific publications;
- The first version of the SEPLN Anthology, including SEPLN articles from 2008 to 2016 and annotated with meta-data and rich linguistic information;
- A faceted-search and visualization interface to explore the SEPLN universe.

The rest of the paper is organized as follows: in Section 2 we describe the scientific text mining tools and resources we have implemented and exploited to extract and process the contents of SEPLN articles so as to build the first version of the SEPLN Anthology corpus. In Section 3 we present the web portal we developed to perform innovative faceted browsing of the SEPLN papers. After briefly reporting related works (Section 4), the paper is concluded with Section 5 where we present our future avenues of research by outlining how to improve the SEPLN Anthology as well as by anticipating the possibilities that

[1] `http://www.sepln.org/`

this resource offers in terms of boosting the involvement of NLP community in scientific text mining by means, for instance, of possible content analysis and visualization challenges.

## 2 Building the SEPLN Anthology

Focusing on the promotion of research in the field of NLP for the past 33 years, the SEPLN has systematically published research articles in both its annual conference proceedings and the SEPLN journal. By allowing the articles to be written in Spanish (as well as in English), SEPLN has also played an important role in fostering NLP research and linguistic technologies targeted to Spanish and in making them accessible to the Spanish-speaking scientific community.

The fact that, over the past few years, each SEPLN article contains titles and abstracts in both Spanish and English constitutes one relevant characteristic not only to boost dissemination, but also to enable new experimentation in different areas of NLP. Both the presence of bi-lingual contents and the peculiar style of SEPLN articles pose specific content extraction and analysis challenges that we address in this work.

In a nutshell, the building of the SEPLN Anthology started by crawling the DBLP Computer Science repository in order to collect the meta-data associated to each research paper (e.g. the *bibtex* entries) and to retrieve the link to the official open access PDF publication for further processing. In the following sections we detail the set of steps we have applied to extract, analyze and enrich the contents of each article in order to support intelligent access to the SEPLN archive. In particular, after a brief overview of our initial dataset in Section 2.1, we describe how we converted SEPLN PDF articles into structured XML documents (Section 2.2). Then we explain how we perform linguistic and semantic analyses of the contents of each article (Section 2.3) and how we enrich these contents by means of metadata retrieved from online web services and knowledge resources (Section 2.4). The textual content of the SEPLN articles of the Anthology, together with the results of their linguistic annotation and enrichment, can be downloaded at the following URL: `http://backingdata.org/seplndata/`.[2]

### 2.1 The SEPLN Journal Dataset

To build the first version of the SEPLN Anthology we collected 505 PDF documents corresponding to the articles published in the SEPLN Journal[3] volumes 40 (2008) to 57 (2016). This dataset includes the following subsets of documents: 374 articles (of about 8 pages), 47 demo articles (2-4 pages), 50 projects' descriptions (about 2-4 pages), 33 PhD thesis' descriptions (about 2-4 pages), and one book review (2 pages).

[2] Annotated papers were saved and made available in the GATE XML data format (`https://gate.ac.uk/sale/tao/splitch5.html##x8-960005.5.2`).

[3] `http://journal.sepln.org`

### 2.2 Converting PDF to XML

Even if the adoption of XML-based formats in scientific publishing is growing considerably, more than 80% of the scientific literature is available as PDF documents. As a consequence, the possibility to consistently extract structured textual content from PDF files constitutes an essential step to bootstrap any scientific text mining process. According to recent evaluations [16, 30] the best performing tools to extract structured contents from PDF articles include:

- GROBID:[4] a Java library that exploits a chain of conditional random field (CRF) sequence taggers to identify the correct semantic class of the textual contents of scientific articles;
- CERMINE:[5] a Java library that relies on both unsupervised and supervised algorithms to spot a rich set of structural features of scientific publications in PDF format;
- PDFX:[6] an online web service that applies a set of layout and content based rules to extract structural textual elements from PDF papers (including title, abstract, bibliographic entries, etc.).

In general, existing tools extract structured contents from PDF articles by generating XML files, sometimes complemented by HTML visualizations of the same content with a different layout. Therefore, when a PDF publication is processed, its original textual layout is completely lost and when the information extracted from a paper is displayed as an HTML page (e.g. to highlight the occurrences of specific words or spot the excerpts presenting the challenges faced by the authors) users are disoriented by the loss of any reference to the layout of the original document.

We addressed this issue by implementing PDFdigest, an innovative application that extracts structured textual contents from scientific articles in PDF format. The output of PDFdigest conversion is a pair of XML and HTML documents. While the XML document is useful to save the structural elements identified in the paper (title, authors, abstract, etc.), the HTML document includes the contents of the paper preserving its original layout. Moreover, each element identified by the markup of the XML document is mapped to the list of identifiers of the DIV elements holding the corresponding textual content in the HTML document. As a result, interactive, layout-preserving HTML-based visualizations of a single paper can be generated, as illustrated in Section 3.

PDFdigest is a Java-based application tailored to deal with both one-column and two-column layouts of PDF articles with textual contents expressed in one or several languages (as SEPLN articles combine Spanish and English).

PDFdigest is able to spot the following core set of structural elements of scientific articles: *title*, *second title* (in case it exists), *authors' names*, *affiliation* and *email*, *abstract(s)*, *categories*, *keyword(s)*, *sections' titles* and *textual content*,

[4] `http://github.com/kermitt2/grobid`
[5] `http://cermine.ceon.pl/`
[6] `http://pdfx.cs.man.ac.uk/`

*acknowledgements* and *bibliographic entries*. PDFdigest can also extract other fields, including *annexes*, *authors' biographies*, *figure and table captions*, *abstract title* and *keywords title*, among others.

The content extraction pipeline implemented by PDFdigest consists of the following five phases: (i) PDF to HTML conversion, (ii) computation of statistics of HTML tags and CSS elements, (iii) rule-based content detection and extraction, (iv) language prediction and, (v) XML generation.

In the first phase–PDF to HTML conversion–we use pdf2htmlEX,[7] obtaining an HTML document that includes DIV elements defining the position and style of small portions of the paper, preserving the paper's original layout by means of CSS properties.

The following phase–computation of structural elements statistics–relies on the JSoup[8] HTML parsing library, which exploits the HTML tags and CSS properties to extract textual contents and identify their semantics (title, abstract, keywords, etc.). Based on this information, PDFdigest computes statistics of the HTML tags and CSS properties used in the document (i.e. the most frequently used font types and size, etc.) and, then, a rule-based extraction phase iterates over the HTML tags and applies a complex set of manually-generated rules to detect and retrieve the structural elements of the paper (i.e. title, abstract, acknowledgements). Specific extraction rules and analysis procedures are implemented for each type of structural element based on the computed layout statistics, the structural markers previously detected, and a set of language-dependent and content-specific regular expressions that can be manually modified or extended.

The final phase performs language prediction of the textual contents of each structural element–to identify, for instance, English and Spanish versions of the title, abstract and keywords–and generates the output file in XML format. A set of language probabilities are calculated individually for each structural element and globally for the whole textual content extracted from each section of the article. The language prediction is computed using the *optimaize language detector*[9] Java API.

In a post-processing step, the generated XML is validated by means of the JTidy[10] and SAX[11] Java parsers.

PDFdigest is highly customizable and can be easily adapted to different styles of scientific articles and languages by modifying the language-dependent regular expressions, the offset thresholds, and (in some special cases) the finite states that define the logical structure of the paper and the extraction and consumption rules themselves.

We evaluated the structured textual content extraction quality of PDFdigest by creating a gold standard set of 30 SEPLN articles manually annotated with

[7] `http://github.com/coolwanglu/pdf2htmlEX`
[8] `http://jsoup.org/`
[9] `http://github.com/optimaize/language-detector`
[10] `http://jtidy.sourceforge.net/`
[11] `http://www.saxproject.org/`

respect to eight types of strctural elements: title(s), abstract(s), list of keywords, section headers (up to a depth of three levels), paragraphs, table captions, figure captions and bibliographic entries. Among all the annotations of structural elements generated by PDFdigest, in our evaluation we considered as true positive items only those annotations spotting the same structural element and covering an identical text span than a gold standard annotation. Over all the structural elements considered, PDFdigest obtained an weighted average F1 score equal to 0.917. The most common content extraction errors, besides tables and figure captions, are due to skipped section titles and bibliographic entries.

### 2.3 Linguistic Analysis

We process the structured textual content extracted from SEPLN Journal Papers in PDF format (see Section 2.2) by means of a customized version of the Dr. Inventor Text Mining Framework [24, 25].[12] We distribute the results of these analyses as textual annotations of the corpus of analyzed SEPLN articles.

We identify sentences in the abstracts and the paragraphs of each paper by means of a sentence splitter customized to scientific publications, thus by properly dealing with expressions like: i.e., et. al., Fig., Tab., etc.

Then we spot tokens inside the textual content extracted from each paper by means of a rule-based language-independent tokenizer developed by relying on ANNIE, the Information Extraction toolbox integrated in GATE [19].

Thanks to a set of JAPE rules[13] [6] customized to the specific citation formats of SEPLN Papers, we identify inline citations in the sentences of each article; then we apply a set of heuristics to link each inline citation to the referenced bibliographic entry. By relying on lexical match rules, we mark inline citations that have a syntactic role in the sentence (e.g. in '*Rossi et al. (2015) discovered that...*', the inline citation '*Rossi et al. (2015)*' is the subject of the sentence).

We exploit the MATE tools[14] [3] to perform both POS-tagging and dependency parsing of publications. Since SEPLN Journal Papers present mixed English and Spanish textual contents, for each text excerpt to analyze (sentence, title, etc.) we rely on the language identified by the PDF-to-text converter to properly select the language-specific POS-tagger and dependency parser to use. When we apply dependency parsing, we consider as a single token only inline citations that have a syntactic role in the sentence where they occur, ignoring the other ones.

In order to make explicit the rhetorical organization of papers, we classify each sentence as belonging to one of the following categories: Approach, Challenge, Background, Outcomes and Future Work.

To this purpose, by exploiting the Weka machine learning platform[15] [33], we trained a logistic regression classifier over the Dr. Inventor Multi-Layer Scientific Corpus [10], a colection of 40 English articles in which each sentence has

[12] `http://driframework.readthedocs.io/`
[13] `http://gate.ac.uk/sale/tao/splitch8.html`
[14] `http://code.google.com/p/mate-tools/`
[15] `http://www.cs.waikato.ac.nz/ml/weka/`

been manually assigned a specific rhetorical category. To enable automated classification, we model each sentence by means of a set of linguistic and semantic features, most of them extracted by relying on the results of the textual analyses described in this Section. We identified the rhetorical category of all the English excerpts of SEPLN Journal Papers,[16] leaving as future work the extension of the classifier to Spanish texts.

By applying a set of 84 JAPE rules formulated and iteratively refined by analyzing a collection of 40 English scientific articles, we spot causal relations inside the English content of SEPLN Papers. Each causal relation is composed of two text excerpts respectively spotting the cause and the related effect.

We perform coreference resolution over the English contents of SEPLN papers by applying a deterministic approach similar to the one proposed by the Stanford Coreference Resolution System[17] [13]. In particular, when building coreference chains we match nominal and pronominal coreferent candidates by considering exact matches, pronominal matches, appositions and predicate nominative.

We support the generation of extractive summaries of SEPLN articles by ranking the sentences of each paper with respect to their relevance to be included in a summary of the same publications. To this purpose we associate to each sentence a summary relevance score: the higher the value of such score is, the more suitable is the sentence to be part of a summary of the paper. In particular, we determine two types of summary relevance scores for each sentence by relying respectively on the TF-IDF similarity of the sentence with the title of the paper [27] and by applying LexRank, a graph-based summarization algorithm [9]. In LexRank we computed the TF-IDF similarity among pairs of sentences by considering the IDF scores of tokens derived from the English and Spanish Wikipedias (2015 dumps). Other methods including centroid and term frequency are implemented based on an available summarization library [27].

### 2.4 Enriching Papers

Besides the linguistic analyses described in the previous Section (2.3), we exploit a set of online resources and web services in order to further enrich the contents of the papers, thus enabling richer navigation and visualization possibilities.

**Mining bibliographic metadata** We retrieve from DBLP the bibliographic metadata of each SEPLN paper, thus collecting the related BibTeX records. To get structured metadata describing the bibliographic entries extracted from each SEPLN article, we rely on three web services:

- Bibsonomy:[18] thanks to the web API of Bibsonomy we are able to retrieve the BibTeX metadata of a bibliographic entry if present in the Bibsonomy database;

[16] All SEPLN Papers have both English and Spanish abstracts and about half of them use English in their body.

[17] `http://nlp.stanford.edu/software/dcoref.shtml`

[18] `http://bitbucket.org/bibsonomy/bibsonomy/wiki/browse/documentation/api/`

– CrossRef:[19] thanks to the bibliographic link web API of CrossRef we can match a wide variety of free-form citations to DOIs if present in CrossRef;
– FreeCite:[20] this online tool analyzes the text of references by relying on a conditional random field sequence tagger trained on the CORA dataset, made of 1,838 manually tagged bibliographic entries[21].

By giving precedence to metadata retrieved from Bibsonomy over CrossRef and Freecite outputs (because of their higher accuracy), it is possible to merge the results retrieved by querying these three REST endpoints, trying to determine for each bibliographic entry the title of the paper, the year of publication, the list of authors, and the venue or journal of publication.

To better characterize the semantics of SEPLN articles, we disambiguate their contents by means of the Babelfy web service[22] [21]. Babelfy spots the occurrences of concepts and Named Entities inside the text of each paper and links them to their right meaning chosen in the sense inventory of Babelnet.

**Author affiliation identification** Our PDF-to-XML conversion approach (introduced in Section 2.2) manages to extract from the header of papers the names and emails of the authors together with the text describing their affiliations. We parse the text of each affiliation by both the Google Geocoding and the DBpedia Spotlight web services in order to try to unambiguously determine the mentioned organization (university, institute, company) together with the city and state where it is located.

– Google Geocoding:[23] useful to identify the name of the institution and its normalized address;
– DBpedia Spotlight:[24] exploited to identify instances of the following types from the DBpedia ontology: Organization, Country and City.

The metadata added to describe each affiliation can be properly merged to try to determine the name and address of the organization referred as well as its geolocation, thus enabling the possibility to explore the geographic dimension of the corpus of SEPLN papers so as to visualize, for instance, the yearly changes of the geographic distribution of the authors who published SEPLN articles.

In addition, we exploit the association of the authors to their email addresses to identify their affiliations when it cannot be done solely based on the information retrieved from the Google and DBPedia services due, for instance, of the presence of multiple institutions in the affiliation text or when there is amibiguity in the identification of the entity that corresponds to the author affiliation.

[19] http://search.crossref.org/help/api/
[20] http://freecite.library.brown.edu/
[21] http://hpi.de/naumann/projects/repeatability/datasets/cora-dataset.html
[22] http://babelfy.org/
[23] http://developers.google.com/maps/documentation/geocoding/intro/
[24] http://demo.dbpedia-spotlight.org/

We evaluated the quality of the automated assignment of institutions to authors by manually checking a randomly-selected subset including 10% of the 1,585 author-paper instances available in the corpus, obtaining a precision value of 0.83 for the unique identification of the affiliations.

## 3 Visualizing SEPLN Articles

The results of SEPLN papers analysis and enrichment are made openly available on a visualization platform accessible at: `http://backingdata.org/sepln/`. For a use case of the dataset and platform the reader is referred to [1].

### 3.1 The Web Visualization Platform

User-friendliness was one of the main goals in the design of the visualization platform. When first accessing it, the user finds a sortable table with basic metadata of the papers of the SEPLN Anthology: title, authors and publication year. On a sidebar, a full-text search box can be used to retrieve documents based on their title, keywords, abstract sentences, author names or affiliations.

Filter fields are populated dynamically with values retrieved from the indexed content, making faceted searches available for an incremental exploration of the corpus. When selected, filters are applied to the search engine calls and immediately reflected in the visualizations offered by the platform. Currently, filters are available for keywords, authors, affiliations, topics, Babelnet synsets, countries, cities and publication years.

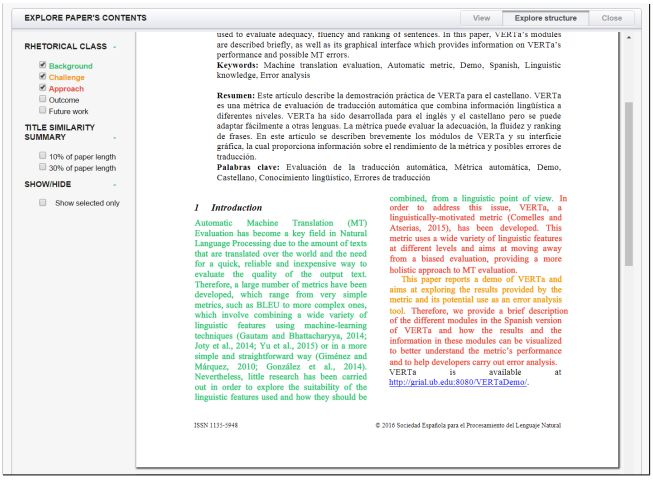


**Fig. 1.** Single-document visualization with rhetorical categories and summaries.

### 3.2 Single-Document Visualizations

The HTML version of the paper produced by PDFdigest is made available with an added layer of interactivity that allows the user to explore the rhetorical categories assigned to the paper's sentences, as well as the sentences identified by the summary ranking algorithms (Fig. 1).

### 3.3 Collection Visualizations

As storage and search solution for the analyzed papers we use Elasticsearch,[25] an open-source application developed in Java and based on the Apache Lucene[26] search engine library. The availability of Elasticsearch clients for the most popular programming languages makes it suitable for our needs as it provides a great flexibility in terms of both indexing and accessing the data. In our project we use version 5.2 of Elasticsearch and its corresponding Java API.[27]

Elasticsearch is distributed together with Kibana,[28] a web-based platform for the analysis and visualization of indexed data. Our web visualization platform takes advantage of Kibana's possibilities for the generation of visualizations but provides a layer on top of it in order to simplify the process of data exploration by means of custom filter and search functionalities.

The following four visualizations currently integrated into our platform showcase this possibility: (i) Date histogram of published papers; (ii) Heatmap of keywords used throughout the years; (iii) Word cloud of keywords; (iv) Heatmap with the evolution of concepts through rhetorical categories over the years.

### 3.4 Geo-spatial Visualizations

Geo-spatial visualization of data makes it possible to better identify relationships for which it can be difficult to get a clear grasp just from tables or graphics. We provide geo-visualizations based on choropleth maps of the world and Spain in which countries (provinces, in the case of Spain) are shaded in proportion to the number of publications available from those countries or provinces, as show in Fig. 2. The different geographic regions are linked by arcs that graphically represent the existence of collaborations among authors affiliated to institutions in the corresponding geographical spaces.

The data used to generate these visualizations is dynamically retrieved from the Elasticsearch index by means of its JavaScript client.[29]

Scalable vector graphics representations of the maps are enriched with data and made responsive to events triggered by the user by means of the D3.js library.[30]

---

[25] `http://www.elastic.co/`
[26] `http://lucene.apache.org/`
[27] `http://elastic.co/guide/en/elasticsearch/client/java-api/`
[28] `http://www.elastic.co/products/kibana/`
[29] `http://elastic.co/guide/en/elasticsearch/client/javascript-api/`
[30] `http://d3js.org/`

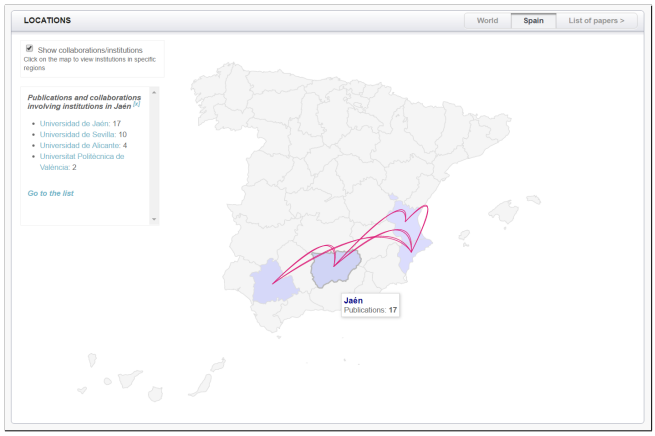


**Fig. 2.** Map-based visualization showing collaborations in Spain.

## 4 Related Work

One of the most comprehensive NLP-related corpora is the ACL Anthology [2] which includes all papers published under the ACL umbrella. The resource was transformed into the ACL Anthology Network (ANN) [23] to enrich the anthology with citation networks information. ANN has been used in different NLP tasks, notably summarization and citation categorization. It was also featured in several exploratory use cases including the Action Science Explorer [8] to investigate the evolution of the field of *dependency parsing* and in [18] to study, among other elements, author collaborations, citations, and funding agencies in the LREC universe. Concerning tools for scientific text exploration and access, in addition to generic scientific document retrieval systems such as Google Scholar or Microsoft Academics, we can mention the ACL Anthology Searchbench system [28] which performs linguistic analysis of the documents and allows the user exploration of the ACL Anthology by means of predicate argument queries and other filter types, and Saffron [20] and Rexplore [22] which are open-domain but perform a reduced set of NLP analyses over publications.

During the last few years tasks related to the extraction, summarization and modeling of information from scientific publications have been proposed in the context of several text mining challenges. Relevant examples are the TAC 2014 Biomedical Summarization Track[31] and the series of Computational Linguistic Scientific Summarization Tasks[32] where the participating teams have been required to generate summaries of scientific articles by taking advantage of the citations they receive. Recently, the participants of the Semantic Publishing Challenges [31] have been required to extract structured information from pa-

[31] http://tac.nist.gov/2014/BiomedSumm/

[32] http://wing.comp.nus.edu.sg/cl-scisumm2017/

pers and model this knowledge as RDF datasets in order to enable the evaluation of the quality of scientific events and publications by means of SPARQL queries. By relying on the data of the Microsoft Academic Graph,[33] several shared tasks have been also proposed dealing with author disambiguation and relevance ranking of papers and institutions [26, 32].

## 5 Conclusions and Future Work

With the availability of massive amounts of scientific publications, research in scientific text mining has proliferated in recent years. However, most approaches perform shallow NLP analysis over scientific contents and consider mainly English publications. In this paper we have described the initial release of the SEPLN Anthology, an automatically analyzed textual corpus created by mining SEPLN publications. We have also introduced a configurable toolkit developed to transform PDF documents into pairs of XML and HTML files that we further analyzed by general and language specific customizable NLP techniques so as to create rich semantic representations of SEPLN documents. We analyze each paper by performing: (i) dependency parsing of English and Spanish sentences, (ii) citation identification, linking, and enrichment, (iii) author information identification, (iv) concept disambiguation, (v) information extraction, and (vi) document summarization. Furthermore, we have also developed a web-based information access platform which exploits the SEPLN Anthology documents to provide interesting single-document and document collection-based visualizations as a means to explore the rich generated contents.

The resources developed in this work are being made available to the research community. We are committed to keep evolving the SEPLN Anthology (e.g., releasing periodic versions, adding functionalities) so as to make it useful in both research and educational activities. There are several avenues of future research we would like to pursue: one of the most relevant is the creation of a gold standard annotation dataset (subset of representative documents) with curated information on authors, rhetorical categories, citations, etc. In parallel, we would like to improve the precision of PDFdigester as well as to perform user-centric evaluations of our web-based interface in order to better understand the value and possibilities of the rich scientific corpora search and browsing patterns we propose.

## Acknowledgements

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the TUNER project (TIN2015-65308-C5-5-R, MINECO / FEDER, UE).

[33] `http://microsoft.com/en-us/research/project/microsoft-academic-graph/`

## References


1. Accuosto, P., Ronzano, F., Ferrés, D., Saggion, H.: Multi-level mining and visualization of scientific text collections. Exploring a bi-lingual scientific repository. In: Proceedings of WOSP 2017 - ACM/IEEE-CS Joint Conference on Digital Libraries. ACM (2017)
2. Bird, S., Dale, R., Dorr, B.J., Gibson, B.R., Joseph, M.T., Kan, M., Lee, D., Powley, B., Radev, D.R., Tan, Y.F.: The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In: Proceedings of LREC (2008)
3. Bohnet, B.: Very high accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd COLING. pp. 89–97. Association for Computational Linguistics (2010)
4. Boudin, F.: TALN archives : une archive numérique francophone des articles de recherche en traitement automatique de la langue. In: TALN 2013. pp. 507–514 (2013)
5. 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)
6. Cunningham, H., Maynard, D., Tablan, V.: JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield (2000)
7. Directorate-General for Research and Innovation (European Commission): Open innovation, open science, open to the world: A vision for Europe (2016)
8. Dunne, C., Shneiderman, B., Gove, R., Klavans, J., Dorr, B.: Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. J. Am. Soc. Inf. Sci. Technol. 63(12), 2351–2369 (2012)
9. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 22, 457–479 (2004)
10. Fisas, B., Ronzano, F., Saggion, H.: A multi-layered annotated corpus of scientific papers. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA) (2016)
11. Holtz, M., Teich, E.: Design of the Darmstadt Scientific Text Corpus (DaSciTex). Technical Report DFG project TE 198/1-1, Technische Universität Darmstadt (2009)
12. Jacsó, P.: Google scholar: the pros and the cons. Online Information Review 29(2), 208–214 (2005)
13. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (pp. 28-34). Association for Computational Linguistics (2011)
14. Ley, M.: DBLP: Some lessons learned. VLDB Endowment 2(2), 1493–1500 (2009)
15. Li, H., Councill, I.G., Lee, W., Giles, C.L.: CiteSeerX: an architecture and web service design for an academic document search engine. In: Proceedings of the 15th WWW Conference. pp. 883–884 (2006)
16. Lipinski, M., Yao, K., Breitinger, C., Beel, J., Gipp, B.: Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 385–386. ACM (2013)
17. Lopez, P.: Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Proceedings of ECDL'09. pp. 473–474. Springer-Verlag, Berlin, Heidelberg (2009)

18. Mariani, J., Paroubek, P., Francopoulo, G., Hamon, O.: Rediscovering 15 + 2 years of discoveries in language resources and evaluation. Language Resources and Evaluation 50(2), 165–220 (2016)
19. Maynard, D., Tablan, T., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., Wilks, Y.: Architectural elements of language engineering robustness. In: Natural Language Engineering. 8(2-3):257–274 (2002)
20. Monaghan, F., Bordea, G., Samp, K., Buitelaar, P.: Exploring your research: Sprinkling some Saffron on semantic Web Dog Food. In: Semantic Web Challenge at the ISWC. vol. 117, pp. 420–435 (2010)
21. Moro, A., Cecconi, F., Navigli, R.: Multilingual word sense disambiguation and entity linking for everybody. In: Proceedings of the 2014 IISWC-PD Conference. pp. 25–28 (2014)
22. Osborne, F., Motta, E., Mulholland, P.: Exploring scholarly data with Rexplore. In: The Semantic Web - ISWC 2013. pp. 460–477 (2013)
23. Radev, D.R., Muthukrishnan, P., Qazvinian, V.: The ACL Anthology Network Corpus. In: NLPIRDL 2009. pp. 54–61 (2009)
24. Ronzano, F., Saggion, H.: Dr. Inventor Framework: Extracting structured information from scientific publications. In: International Conference on Discovery Science 2015. Springer, pages 209–220 (2015)
25. Ronzano, F., Saggion, H. Knowledge extraction and modeling from scientific publications. In the Proceedings of the Workshop Semantics, Analytics, Visualisation: Enhancing Scholarly Data co-located with the 25th International World Wide Web Conference. (2016)
26. Roy, S. B., De Cock, M., Mandava, V., Savanna, S., Dalessandro, B., Perlich, C., Hamner, B.: The microsoft academic search dataset and kdd cup 2013. In Proceedings of the 2013 KDD cup 2013 workshop. p. 1. (2013)
27. Saggion, H.: SUMMA: A robust and adaptable summarization tool. Traitement Automatique des Langues 49(2). pp103–125 (2008)
28. Schäfer, U., Kiefer, B., Spurk, C., Steffen, J., Wang, R.: The acl anthology searchbench. In: Proceedings of the 49th ACL. pp. 7–13 (2011)
29. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.P., Wang, K.: An overview of Microsoft Academic Service (MAS) and applications. In: Proceedings of the 24th WWW Conference. pp. 243–246 (2015)
30. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P.J., Bolikowski, Ł.: CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR) 18(4), 317–335 (2015)
31. Vahdati, S., Dimou, A., Lange, C., Di Iorio, A.: Semantic publishing challenge: bootstrapping a value chain for scientific data. In International Workshop on Semantic, Analytics, Visualization. pp. 73–89 (2016)
32. Wade, A. D., Wang, K., Sun, Y., Gulli, A.: WSDM Cup 2016: Entity Ranking Challenge. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 593–594 (2016)
33. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2016)