A curation pipeline and web-services for PDF documents

André Santos[1], Sérgio Matos[1], David Campos[2] and José Luís Oliveira[1]

[1]DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
{aleixomatos,andre.jeronimo,jlo}@ua.pt
[2]BMD Software, 3810-074 Aveiro, Portugal
david.campos@bmd-software.com

Abstract

The continuous growth of the biomedical literature and the need to efficiently find and extract information from its content led to the development of various text mining tools. More recently, these tools started being integrated in user-friendly applications facilitating their use by expert database curators. However, these tools were mainly designed to extract information from text based documents, in XML and other formats, while today a considerable part of the biomedical literature is published and distributed in PDF format.

To address this limitation, we extended the web-based literature curation tool Egas, adding support for direct document curation and annotation over PDF files, with side-by-side visualization of the original PDF document and of the extracted textual content. Egas' PDF document processing and text-mining features are supported by a newly developed web-services platform built over Neji, a highly efficient information extraction framework. These web services allow integrating PDF text extraction and annotation capabilities to other tools and text mining pipelines.

1 Introduction

The large amount of information and knowledge continuously produced in the biomedical domain is reflected on the number of published journal articles. In 2015, the bibliographic database MEDLINE contained over 23 million references to journal articles in life sciences, of which 1 million were added in that year (U.S. National Library of Medicine, 2016). At this rate, staying updated with the current knowledge and identifying the most relevant publications and information on a given subject is a very challenging task for researchers.

To facilitate the access to knowledge, several resources started by manually curating scientific articles, extracting and structuring relevant and validated information. However, with the rapid growth of data this task became unfeasible (Yeh et al., 2003; Rebholz-Schuhmann et al., 2005), and automatic information extraction tools were developed and integrated in the curation pipeline in order to accelerate the curation process (Neves and Leser, 2012). This also led to the need of creating end-user interfaces to these tools, allowing their use by curators in a efficient manner. The success of the BioCreative Interactive Annotation Task series demonstrates the importance of these efforts (Arighi et al., 2013).

While existing information extraction tools have been shown to achieve robust performance in various tasks, and various literature curation tools have been proposed that make use of such automated methods, they were generally designed to work with plain text or with structured formats such as XML. There is however a lack of tools for supporting curation workflows that make use of the Portable Document Format (PDF), which has become one of the most popular file formats for publishing and sharing documents.

We have previously presented Neji (Campos et al., 2013), an open source framework for biomedical concept recognition, and Egas (Campos et al., 2014), a web-based tool for literature curation built with modern web technologies and providing simple inline representation of annotations and user-friendly interaction. In this paper we present new features added to Egas and Neji to support text-mining and curation workflows over PDF documents. In Section 2 we describe Neji's new PDF processing functionalities and present its

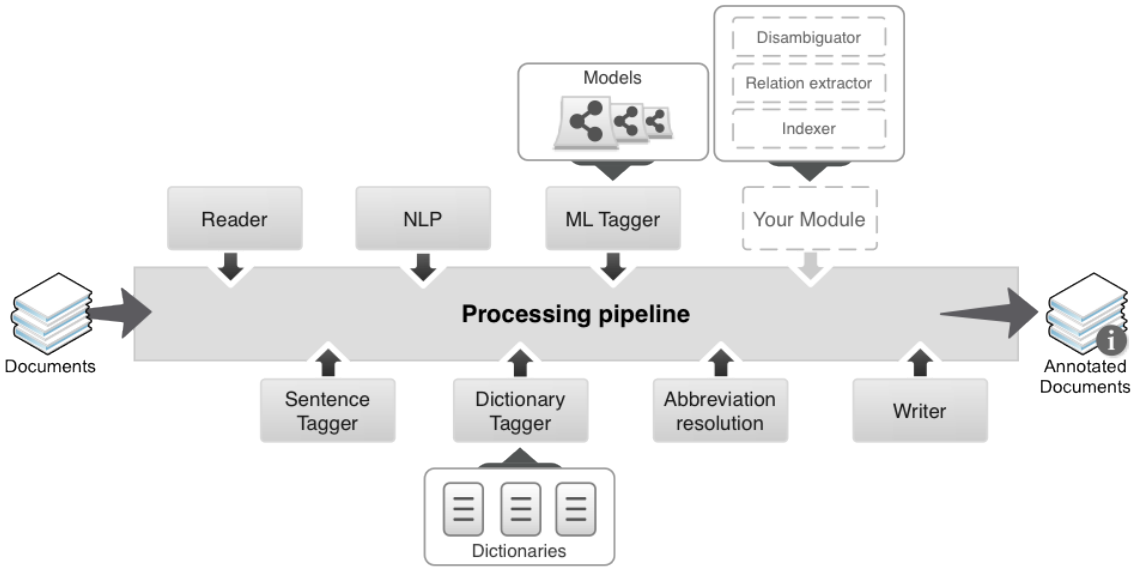

Figure 1: Neji processing pipeline and modular architecture (Campos et al., 2013)

new web-services platform. These web-services are used by the curation tool for extracting the text from PDF documents and for obtaining automatic concept annotations, and also facilitate the integration of Neji's functionalities in external text-mining pipelines and tools. Egas is described in Section 3, highlighting the new PDF annotation features including side-by-side synchronous visualization of the extracted text and of the original PDF, and also the display of concept annotations over the PDF document.

2 Neji

Neji is an open source framework for biomedical concept recognition built around four crucial characteristics: modularity, scalability, speed and usability. It follows several state-of-the-art methods for biomedical natural language processing (NLP), namely methods for sentence splitting, tokenization, lemmatization, POS, chunking and dependency parsing. The concept recognition tasks are performed using dictionary matching and machine learning techniques with normalization. This framework implements a very flexible and efficient concept tree to store the document annotations, supporting nested and intersected concepts with one or more identifiers. It supports several input and output formats including the most popular ones in biomedical text mining, such as IeXML, Pubmed XML, A1, CONLL and BioC. The architecture of Neji allows users to configure the processing of documents according to their specific objectives and goals, for example by simply combining existing or new modules for reading, processing and writing data, or by selecting the appropriate dictionaries or machine learning models according to the concept types of interest.

Neji has been evaluated on several corpora, covering different concept types (Campos et al., 2013; Campos et al., 2015; Matos et al., 2016). Table 1 shows a summary of the concept identification performance.

2.1 Pipeline and modules

The main component of Neji is the processing pipeline (Figure 1), a series of independent modules, each of them responsible for a specific processing task, that are executed sequentially. We used Monq.jfa[1], a library for fast and flexible text filtering with regular expressions, to implement each pipeline module as a custom deterministic finite automaton (DFA) with specific rules and actions.

2.1.1 Handling PDF files

Thanks to Neji's modular architecture, adding PDF processing capabilities only required the implementation of a new reader module. For this, we integrated LA-PDFText (Ramakrishnan et al., 2012), a state-of-the-art open-source tool for handling PDF documents. LA-PDFText makes use of a carefully crafted set of rules defined on the business rules management system DROOLS, allow-

[1] http://www.pifpafpuf.de/Monq.jfa/

Table 1: Neji concept recognition results on a variety of corpora and concept types. D: Dictionary; ML: Machine-Learning

Corpus	Concept type	F-score	Method
CRAFT	Species	95%	D
	Cell	92%	D
	Gene and Protein	76%	ML
	Chemicals	65%	D
	Cellular Component	83%	D
	Biological Process and Molecular Function	63%	D
NCBI Disease	Disorders	85%	D
Anem	Anatomy	82%	D
BC II Gene Mention	Gene and Protein	87%	ML
tmVar	Genetic Variants	86%	ML
BC IV ChemdNER	Chemicals	87%	ML

ing to correctly handle different PDF layouts such as one column, two columns and mixed layouts. This feature also allows defining different sets of rules for specific PDF layouts if necessary, and we therefore included in the new Neji reader an optional parameter for this.

In order to evaluate the text extraction quality, we obtained the original PDF documents corresponding to the 67 full-text articles that compose the CRAFT corpus (Bada et al., 2012), and compared the text extracted by LA-PDFText, through our processing pipeline, to the distributed text contents, which were extracted from XML files. For these articles, published in 21 different journals and having distinct layouts, we obtained an exact match in 90% of the extracted sentences.

Apart from extracting the text, which is sufficient for running the processing pipeline, we added additional capabilities to the reader, in order to make use of PDF processing in the curation tool Egas. Namely, we apply sentence splitting to the extracted chunks of text, and extract the position of each sentence in each page to allow aligning and navigating between the plain text and PDF views in the user interface. This information is associated to each sentence and carried over to the remaining modules in the pipeline. A new writer module was also implemented that exports this extended information in JSON format, for simple reuse in external tools.

2.2 Web-services

Neji web-services are intended to facilitate the use and access to Neji functionalities by providing a simple RESTful API that allows developers to send their input documents and receive the plain text extracted from the submitted PDF file and also annotation results in various well-known formats, including standoff (A1) (Kim et al., 2009; Stenetorp et al., 2012) and BioC (Comeau et al., 2013).

Different annotation services can be configured in the platform, in which a service is an annotation pipeline with a custom set of resources (dictionaries and ML models) and processing properties. This provides a way to easily manage concurrent annotation services, allowing the configuration of the properties and resources of each of them independently. Additionally, resources are loaded into memory as soon as a new service is created. Since this usually is an expensive step, especially for large ML models, having the resources in memory greatly reduces the total annotation time.

3 Egas

Egas is a web-based platform for biomedical text mining and collaborative curation that supports inline annotation of concept occurrences and of relations between these concepts. Annotations can be performed automatically, using the available services for automatic concept and relation identification, or manually, wherein a user can add new annotations and also edit or remove automatically generated annotations. The results can be then exported to various standard annotation formats.

To adapt Egas to support literature curation over PDF documents, we integrated PDF text extraction using the Neji web services RESTful API and adapted the interface for side-by-side visualization of the extracted text alongside the original PDF

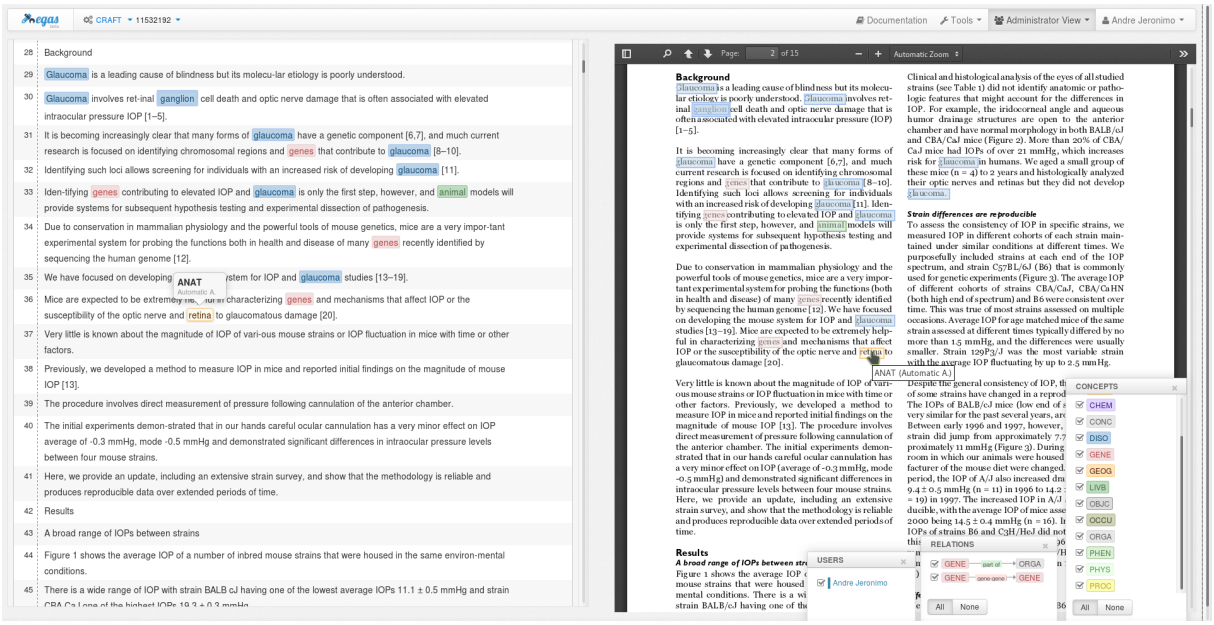

Figure 2: Egas PDF annotation interface

document, allowing the navigation between both zones, synchronizing the text annotation area and the PDF visualization area.

Egas' file import web services were also extended to support PDF files. As with the remaining file formats, this web service is responsible for receiving the file, extracting the text content using Neji's PDF processing feature as described above, and creating the whole data structure to support document annotations. This structure includes also sentence information retrieved from Neji, such as the start and end indexes, with respect to the extracted plain text, and its position within the PDF page, allowing synchronous scrolling and navigation between the plain text and PDF views.

Figure 2 shows Egas' user interface for PDF annotation. The original PDF document is displayed on the right-side panel, while the left panel shows the annotation panel with the extracted text, allowing annotation using the same simple interactions as for other document formats, as described in (Campos et al., 2014). As can be seen in the figure, concept annotations added by the automatic annotation services or by the curator are displayed on the plain text as well as on the PDF document. Additionally, a tooltip with information associated to each annotation is shown when hovering the mouse over the annotation on either panel. By clicking a sentence number on the annotation panel, the PDF document is scrolled accordingly, and the corresponding sentence is briefly highlighted to facilitate its identification. Conversely, double-clicking a sentence on the PDF scrolls the text on the annotation panel and highlights the corresponding sentence.

4 Conclusions

Assisted literature curation tools, based on text mining and information extraction methods, are increasingly being used by curation teams, helping to expedite their tasks. However, there is a lack of tools that support direct annotation of PDF documents, which is a very common format for the scientific literature and other document types, such as patents. We present a new feature of Egas that allows direct document curation and annotation over PDF files, with side-by-side visualization of the original PDF document and of the extracted textual content. By aligning the user-friendliness of Egas with the possibility of reading the document in a very familiar format such as PDF, we provide a more convenient and agreeable literature curation environment, which could contribute to improved efficiency.

References

Cecilia N Arighi, Ben Carterette, K Bretonnel Cohen, Martin Krallinger, W John Wilbur, Petra Fey, Robert Dodson, Laurel Cooper, Ceri E Van Slyke, Wasila Dahdul, et al. 2013. An overview of the biocre-

ative 2012 workshop track iii: interactive text mining task. *Database*, 2013:bas056.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the craft corpus. *BMC bioinformatics*, 13(1):1.

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. A modular framework for biomedical concept recognition. *BMC bioinformatics*, 14(1):281, jan.

David Campos, Jóni Lourenço, Sérgio Matos, and José Luís Oliveira. 2014. Egas: a collaborative and interactive document curation platform. *Database : the journal of biological databases and curation*, 2014, jan.

David Campos, Sérgio Matos, and José L Oliveira. 2015. A document processing pipeline for annotating chemical entities in scientific documents. *Journal of cheminformatics*, 7(1):1.

Donald C Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, et al. 2013. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013:bat064.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics.

Sérgio Matos, David Campos, Renato Pinho, Raquel M Silva, Matthew Mort, David N Cooper, and José Luís Oliveira. 2016. Mining clinical attributes of genomic variants through assisted literature curation in egas. *Database*, 2016:baw096.

Mariana Neves and Ulf Leser. 2012. A survey on annotation tools for the biomedical literature. *Briefings in bioinformatics*, page bbs084.

Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, and Gully Apc Burns. 2012. Layout-aware text extraction from full-text PDF of scientific articles. *Source code for biology and medicine*, 7(1):7, jan.

Dietrich Rebholz-Schuhmann, Harald Kirsch, and Francisco Couto. 2005. Facts from textis text mining ready to deliver? *PLoS Biol*, 3(2):e65.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. pages 102–107, apr.

U.S. National Library of Medicine. 2016. Detailed Indexing Statistics: 1965-2015.

Alexander S Yeh, Lynette Hirschman, and Alexander A Morgan. 2003. Evaluation of text data mining for database curation: lessons learned from the kdd challenge cup. *Bioinformatics*, 19(suppl 1):i331–i339.