Development of Text Mining Tools for Information Retrieval from Patents

Tiago Alves[1,2(✉)], Rúben Rodrigues[1], Hugo Costa[2], and Miguel Rocha[1]

[1] Centre Biological Engineering, University of Minho, 4710-057 Braga, Portugal
tiago_alves26@hotmail.com
[2] Silicolife Lda, 4715-387 Braga, Portugal

Abstract. Biomedical literature is composed of an ever increasing number of publications in natural language. Patents are a relevant fraction of those, being important sources of information due to all the curated data from the granting process. However, their unstructured data turns the search of information a challenging task. To surpass that, Biomedical text mining (BioTM) creates methodologies to search and structure that data. Several BioTM techniques can be applied to patents. From those, Information Retrieval is the process where relevant data is obtained from collections of documents. In this work, a patent pipeline was developed and integrated into @Note2, an open-source computational framework for BioTM. This integration allows to run further BioTM tools over the patent documents, including Information Extraction processes as Named Entity Recognition or Relation Extraction.

Keywords: Biomedical text mining · Patents · Information retrieval task · PDF to text conversion · @Note2

1 Introduction

Huge amounts of information are generated every day. In the life sciences, the number of publications, reports and patents available on databases is increasing considerably [1,2]. Patents are validated documents representing the intellectual property rights of an invention, being important sources of information due to their novelty nature, with exclusive data that is not published in other scientific literature [3,4]. So, exploring them is critical to understand several biological fields [3,5]. However, the access to these documents is limited. There are some systems able to extract some patent sections. This is the case with SureChEMBL, a tool that searches for chemicals and their structure on patents [6].

Patent documents are available in numerous databases. Those which have grant protection only for specific countries can be used for localized searches. For general-purpose searches, worldwide databases with patents with international protection are a more viable option. The j-PlatPat from Japan Patent Office (JPO) or PatFT from the United States Patent and Trademark Office (USPTO) are databases included in the former group, while the PATENTSCOPE from

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_9

World Intellectual Property Organization (WIPO) or esp@cenet from European Patent Office (EPO) are included in the latter [4].

For instance, the WIPO database has 2.7 million patents registered only in 2014 [7–9]. Since these large amounts of data are available in an unstructured nature without annotations about the text structure and available entities, the search and extraction of relevant information is a difficult and time-consuming task, impossible to be done manually [7]. To exploit these data, automating that process, the Biomedical Text Mining (BioTM) field emerged [10]. It is based on different knowledge areas such as statistics, artificial intelligence or management science, combined with text analytic components as Information Retrieval (IR), Information Extraction (IE) or Natural Language Processing (NLP) [11]. From these, IR allows to obtain relevant information resources (e.g. papers or patents) from an extensive collection of documents, and IE allows the extraction of pertinent information from these documents [12].

To apply BioTM techniques, text files are usually the input. However, patent documents are typically accessed in Portable Document Format (PDF) files, coming from encrypted image files, usually BMP, TIFF, PNG or GIF. So, the conversion of these files into machine-coded, readable, editable and searchable data is mandatory. For that, methods as Optical Character Recognition (OCR) are used [13]. The process can be summarized in two main processes: character extraction, where learned patterns are applied to delimit words or individual letters; and character recognition, where words are identified [14].

Several BioTM platforms has been developed by the scientific community. @Note2[1], developed by the University of Minho and the SilicoLife company is among these efforts. As a Java multi-platform BioTM Workbench, @Note2 uses a relational database and is based on a plug-in architecture, allowing the development of new tools/methodologies in the BioTM field [15].

Structurally, @Note2 is organized into core libraries and user interface tools. The core libraries are organized in three main functional modules: the Publication Manager Module (PMM), which can search documents on online repositories (IR Search process) and download their respective full-text documents (IR Crawling process); the Corpora Module (CM), responsible for corpora management, creating and applying IE processes to them with a manual curation system; and the Resources Module (RM), which allows the management of lexical resources to be used in IE processes. The user interface tools allow a simples interaction with the user to configure and use @Note2's functionalities [15].

Here, the objective was to develop a pipeline, a new plug-in to @Note2, able to make patent data amenable to be searched and used as an information source for the IE processes already available in @Note2 and BioTM in general.

2 Patent Pipeline Development

The patent pipeline can be organized into four different tasks. It can search for patent IDs, retrieve patent metadata, download the published patent PDF file,

[1] http://anote-project.org/.

and, finally, apply PDF to text conversion methodologies to those files. Each task was structured into a module with specific inputs and outputs. Thus, sources to search and retrieve patent IDs, to search for metadata about each patent and to return the patent file(s) in PDF format were configured as components of the *search sources module*, *metainformation sources module* and *retrieval sources module*, respectively. The used PDF to text conversion methodologies were organized in the *PDF conversion module* (Fig. 1).

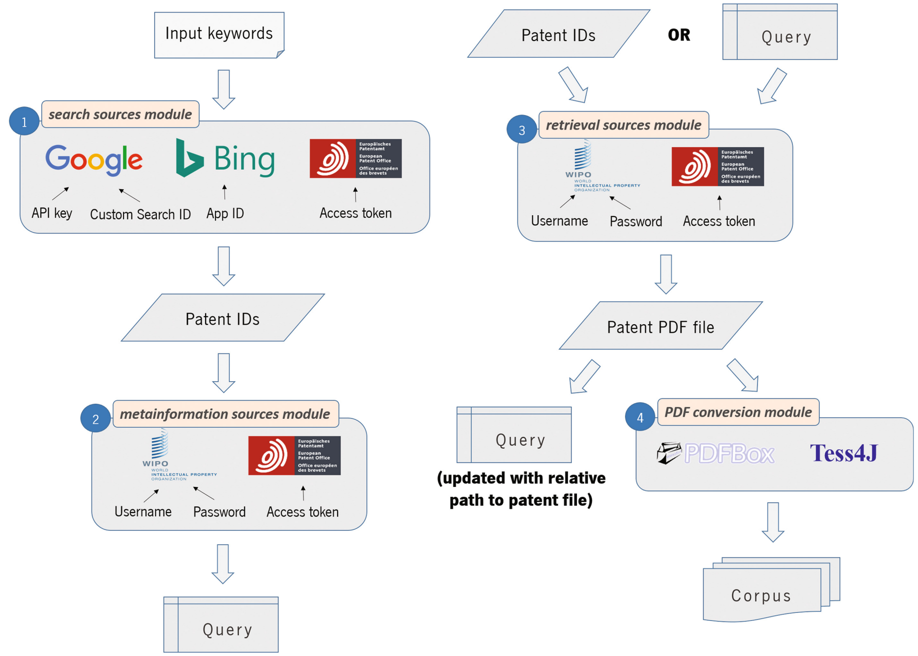
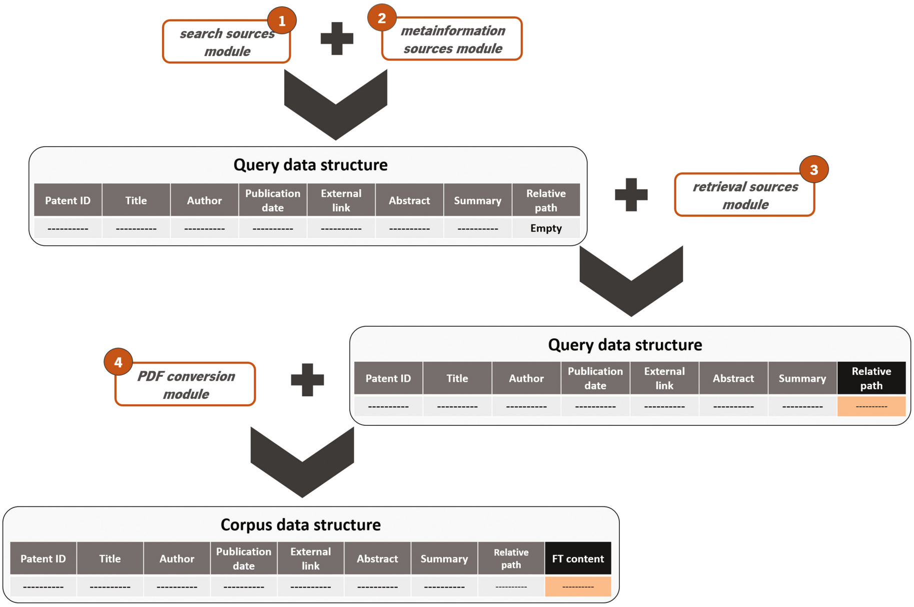


Fig. 1. Summary of the designed patent pipeline (numbers represent the process flow).

To get any result using the first three modules, specific access keys resulting from the services registration are required to get access to servers and retrieve the requested data. To start the search process, input keyword(s) are required, which may be biomedical entities as chemicals, genes, diseases, among others. These keywords are then processed by the *search sources module*. Into this module, two popular search engines (the *Custom Search* API from Google and the *Bing Search* API from Microsoft) and the *Open Patent Services (OPS) web services* API from EPO were used. The two first were configured to retrieve patent IDs from Google Patents, with around 87 millions of patents from 17 countries [16]. The result is the union of the patent IDs returned by all components.

The *metainformation sources module* returns the invention title, authors, publication date, a link to a patent database entry (if available) and the abstract to each patent. When available, the description and claims are also extracted.

To avoid repetitions, the patent family is extracted and only one ID is used to retrieve metadata, being the others saved as external references. That data is then stored into *query*, a data structure from @Note2 to save the document information (Fig. 2). Two different services were configured: the *PATENTSCOPE web service* API from WIPO and the *OPS web service* API from EPO.



Query data structure

Patent ID	Title	Author	Publication date	External link	Abstract	Summary	Relative path
----------	----------	----------	----------	----------	----------	----------	Empty

Query data structure

Patent ID	Title	Author	Publication date	External link	Abstract	Summary	Relative path
----------	----------	----------	----------	----------	----------	----------	----------

Corpus data structure

Patent ID	Title	Author	Publication date	External link	Abstract	Summary	Relative path	FT content
----------	----------	----------	----------	----------	----------	----------	----------	----------

Fig. 2. Creation and update process for *query* and *corpus* data structures. The numbers represent the modules of the pipeline and their flow. The orange *query* data field represents the update process of the original *query*, while the orange *corpus* data field represents the field that turns the *corpus* into a different data structure.

The *retrieval sources module* returns the patent PDF files, saving their path into the *query* (Fig. 2). This module uses the same APIs from the previous with different configurations. Both metainformation and PDF retrieval modules use a sequential architecture. The first takes all the patents, while the next components receive only the ones that did not get any result. That process is repeated until all patents are processed or all components were used.

The *PDF conversion module* takes all the files from the previous module, extracting their text. As shown in Fig. 2, this allows the creation of a *corpus*, allowing to run IE methods, for instance, NER or RE. In this module, alongside with *Apache PDFBox* library (already implemented on @Note2) it was configured the *Tess4J*, version 3.2.1 (developed by Quan Nguyen) implementing *Tesseract*, an OCR algorithm from Google, and also a hybrid method combining these two methodologies. The *Apache PDFBox* allows to extract the Unicode text available on PDF documents. The hybrid method allows a previous PDF treatment, improving their quality to be processed by *Tess4J* system.

On @Note2, patent handling features were inserted in different core libraries. The patent ID search and metadata retrieval were added as new IR Search processes called "Patent Search", while the patent PDF file download was added as a new IR Crawling process, and the new PDF to text conversion methods were put into the Corpora Module as a pre-processing method (Fig. 3).

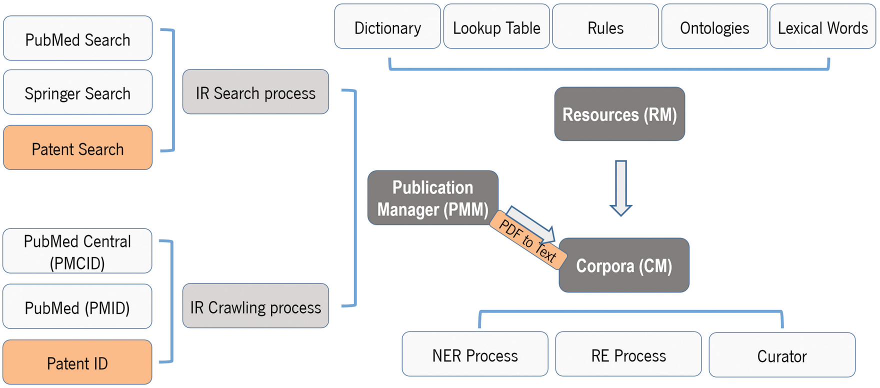


Fig. 3. @Note2 structure with patent pipeline implementations. The orange boxes represent the new components added.

3 Results

The pipeline is materialized by a plug-in allowing patent search in Google Patents and esp@cenet repositories. A graphical interface was made to set @Note2 Preferences where credentials can be saved. The main wizard includes two steps (Fig. 4): the keywords and the *query* name input pane; and the configurations pane, where the previous defined configurations can be edited (Fig. 5).

To test the system, data from the 1000 patents with the longest abstracts from the BioCreative V CHEMDNER task were used (IDs, titles and abstracts). The abstract was tokenized and compared with the tokens from our PDF to text conversion. In this comparison, we used the Smith-Waterman algorithm, a Dynamic Programming algorithm to evaluate the matches. This allows calculating performance metrics as precision, recall and F1 values (based on the number of tokens that match exactly on the texts). Alongside the accuracy calculation, it is possible infer the amount of conversion errors, as well as verify the number of documents correctly downloaded.

Complete metadata were extracted for 917 patents (91,7%). From the remaining 83, 76 were filled partially. Then, also 993 patent PDF files were correctly obtained (99,3%). For both processes, the success rate was limited due to repositories coverage and to restrictions imposed by the use of free credentials.

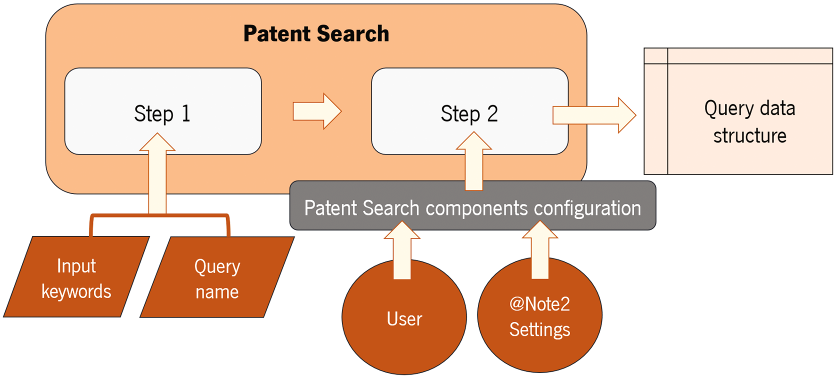
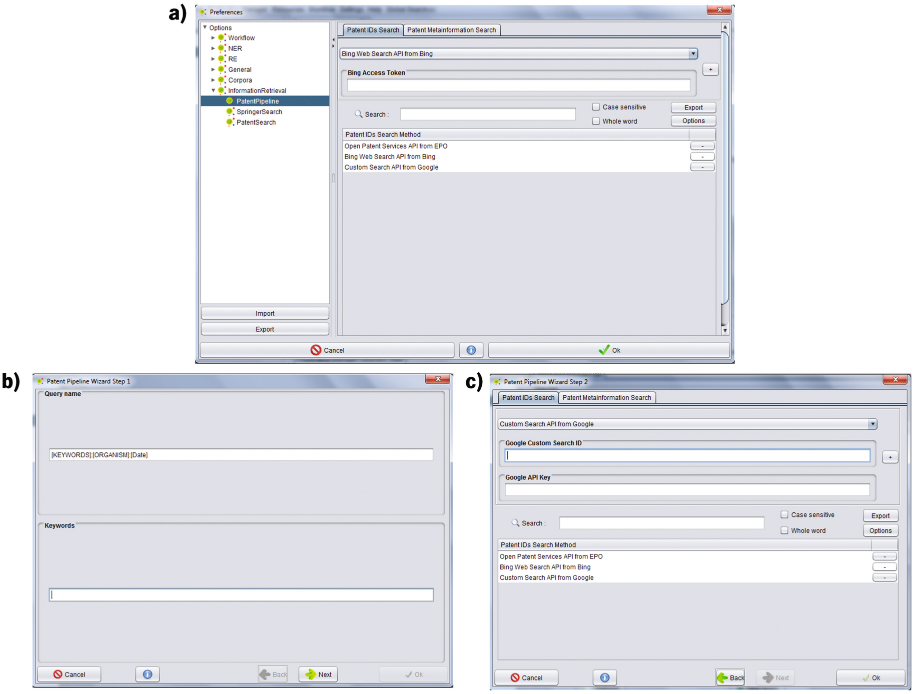


Fig. 4. @Note2 Patent Search plug-in. The pipeline uses input keywords, the *query* name and configurations provided by the user or by @Note2 settings to search for patent IDs and to download patent metadata.



a)

Bing Web Search API from Bing

Bing Access Token:

□ Case sensitive
□ Whole word

Patent IDs Search Method
Open Patent Services API from EPO
Bing Web Search API from Bing
Custom Search API from Google

b)

Query name: [KEYWORDS]:[ORGANISM]:[Date]

Keywords:

c)

Custom Search API from Google

Google Custom Search ID:

Google API Key:

□ Case sensitive
□ Whole word

Patent IDs Search Method
Open Patent Services API from EPO
Bing Web Search API from Bing
Custom Search API from Google

Fig. 5. @Note2 Patent Search GUI. (a) panel for @Note2 preferences; (b) and (c) Steps 1 and 2 from the Patent Search Wizard, respectively.

From the PDF to text evaluation (Fig. 6), the precision values showed a small variance being high in all documents (mean around 95%), while recall values were higher than 80% for 75% of all documents. However, 94 documents returned a recall value under 10% representing old patents (some patents before the 1970s) with only some drawings and a brief description, being the full text data absent. As expected, this led to a high standard deviation (around 30%) which can be also explained by the presence of a high number of chemical structures or formulas that are omitted in the BioCreative task abstract text or simply are converted to noise. The F1 measure summarizes the system capacity to transform most of the PDF files into readable text. Since some patent files have more than 200 pages, to process 1000 patents, the whole pipeline took around 3 days using a PC with an i7 960 @ 3.2 GHz processor and 16 GB of RAM.

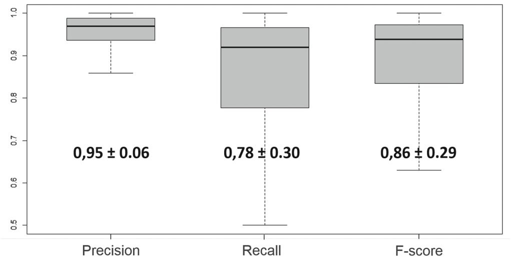


Fig. 6. Boxplots for the evaluation metrics of the PDF to text conversion process. The mean and standard deviation are given in bold.

4 Conclusions

Recently, patents have been a target for BioTM techniques since they are a great source of information for many fields. Based on @Note2, IR Search and Crawling processes were designed and implemented, allowing the search and retrieval of patent information and respective documents. Also, new improvements were made to the @Note2 PDF to text conversion system. Testing these processes with a set of 1000 patents from a BioCreative V task shows that nearly all PDFs were correctly downloaded with respective metadata. Using the new PDF to text system on that documents, we got around 85% of F-score.

The main innovation of this work was the creation of new IR processes applied to patents surpassing common problems related to searching and retrieving those documents, allowing also the posterior implementation of several IE techniques to those texts. Since @Note2 is an open-source software, this framework opens doors to the community to take advantage of all sections from the published patents with biological relevance more easily and without the need to expend large amounts of time browsing several databases. To @Note2, the integration of these tools allows developing an extensive set of text mining pipelines over patents, which were only possible for scientific articles so far.

Some improvements can still be made, namely reducing the processing time and adding new components in each module using the designed architecture.

Acknowledgments. This work is co-funded by the North Portugal Regional Operational Programme, under “Portugal 2020”, through the European Regional Development Fund (ERDF), within project SISBI- Ref^aNORTE-01-0247-FEDER-003381. This study was also supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte (NORTE-01-0145-FEDER-000004) funded by European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte.

References

1. Faro, A., Giordano, D., Spampinato, C.: Combining literature text mining with microarray data: advances for system biology modeling. Brief Bioinform. **13**(1), 61–82 (2012)
2. Klinger, R., Kolarik, C., Fluck, J., Hofmann-Apitius, M., Friedrich, C.M.: Detection of IUPAC and IUPAC-like chemical names. Bioinformatics **24**(13), i268–i276 (2008)
3. WIPO, Guidelines for Preparing Patent Landscape Reports (2015)
4. Latimer, M.T.: Patenting inventions arising from biological research. Genome Biol. **6**(1), 203 (2005)
5. WIPO, WIPO Guide to Using Patent Information (2015)
6. Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S.A., Pettersson, J., Goncharoff, N., Hersey, A., Overington, J.P.: Surechembl: a large-scale, chemically annotated patent document database. Nucleic Acids Res. **44**(D1), D1220–D1228 (2016)
7. Wu, C., Schwartz, J.M., Brabant, G., Peng, S.L., Nenadic, G.: Constructing a molecular interaction network for thyroid cancer via large-scale text mining of gene and pathway events. BMC Syst. Biol. **9**(Suppl. 6), S5 (2015)
8. Lu, Z.: Pubmed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford), vol. 2011, p. baq036 (2011)
9. WIPO, World Intellectual Property Indicators, 2015th edn. World Intellectual Property Organization - Economics and Statistics Division (2015)
10. Cohen, K.B., Hunter, L.: Getting started in text mining. PLoS Comput. Biol. **4**(1), e20 (2008)
11. Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., Fast, A.: Practical text mining and statistical analysis for non-structured text data applications. Academic Press (2012)
12. Krallinger, M., Valencia, A.: Text-mining and information-retrieval services for molecular biology. Genome Biol. **6**(7), 224 (2005)
13. Asif, A.M.A.M., Hannan, S.A., Perwej, Y., Vithalrao, M.A.: An overview and applications of optical character recognition. Int. J. Adv. Res. Sci. Eng. **3**(7) (2014)
14. Holley, R.: How good can it get? analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine **15** (2009)
15. Lourenço, A., Carreira, R., Carneiro, S., Maia, P., Glez-Peña, D., Fdez-Riverola, F., Ferreira, E.C., Rocha, I., Rocha, M.: @note: a workbench for biomedical text mining. J. Biomed. Inform. **42**(4), 710–720 (2009)
16. Google, About google patents (2017)