NLEL-MAAT at CLEF-IP

Santiago Correa, Davide Buscaldi, Paolo Rosso.
NLE Lab, ELiRF Research Group, DSIC,
Universidad Politécnica de Valencia, Spain.
{scorrea, dbuscaldi, prosso}@dsic.upv.es
http://users.dsic.upv.es/grupos/nle

Abstract. This report presents the work carried out at NLE Lab for the IP@CLEF-2009 competition. We adapted the JIRS passage retrieval system for this task, with the objective to exploit the stylistic characteristics of the patents. Since JIRS was developed for the Question Answering task and this is the first time its model was used to compare entire documents, we had to carry out some transformations on the patent documents. The obtained results are not good and show that the modifications adopted in order to use JIRS represented a wrong choice, compromising the performance of the retrieval system.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval;

General Terms

Measurement, Performance, Experimentation

Keywords

Passage Retrieval, Intellectual Property.

1 Introduction

The IP@CLEF-2009 arises from the growing interest by different business and academy sectors in the field of Intellectual Property (IP). The task consists in finding patent documents that constitute prior art to a given patent. Passage Retrieval (PR) systems are aimed at finding parts of text that present a high density of relevant information [1]. We based our work on the assumption that the density of the information in patent documents is high enough to be exploited by means of a PR system. Therefore, we adapted the JIRS PR system to work on IP-CLEF data.

JIRS[1] is an open source PR system which was developed at the *Universidad Politécnica de Valencia* (UPV), primarily for the Question Answering (QA) task. It ranks passages depending on the number, length and positions of the query *n*-grams that are found in the retrieved passages. In our previous participations to the QA task, JIRS proved to be superior in PR performance to the Lucene[2] open source system [2]. In the following sections, we explain the main concepts of JIRS system and we discuss how it has been applied in solving the problem; in Section 5 we discuss the obtained results.

2 Intellectual Property Task

The main task consists in finding the prior art for a given patent. The corpus is composed by documents from the *European Patent Organization* (EPO)[3] published between 1985 and 2000, a total of 1,958,955 patent documents

[1] http://sourceforge.net/projects/jirs/
[2] http://lucene.apache.org
[3] http://www.epo.org/

relating to 1.022.388 patents. The provided documents are encoded in *XML* format, emphasizing these sections: title, language, summary and description.

A total of 500 patents are analyzed using the supplied corpus to determine their prior art; for each one of them the systems must return a list of 1000 documents with their score ranking.

3 The passage retrieval engine JIRS

The passage retrieval system JIRS is a based on *n*-grams (an *n*-gram is a sequence of *n* adjacent words). JIRS has the ability to find structures of words sequences in a large collection of documents quickly and efficiently, through the use of different n-grams models. In order to do this, JIRS searches for all possible n-grams of the words sequence in the collection and it gives them a weight in relation to the n-grams quantity and weight that appear in these passages. E.g.: suppose to search a document collection, using the JIRS system, in order to find articles related to the phrase: “anti-lock braking system”; The system could retrieve the following two passages: “…braking system consists of disk brakes …” and “…anti-lock braking system developed by…”. In a standard IR engine the first passage would obtain a higher weight due to the repetitions of the words with the “brak” stem. In JIRS the second passage is ranked higher because of the presence of the 3-gram “anti-lock braking system”. In order to calculate the n-grams weight of each passage, first of all it is necessary to identify the most relevant n-gram and assign to it a weight equal to the sum of all term weights. The weight of each term is set to:

$$w_k = 1 - \frac{\log(n_k)}{1+\log(N)} \tag{1}$$

Where n_k is the number of passages in which the term appears and N is the total number of passages in the system.

The target is to establish a measure of similarity between a passage (d) and a text (q).

$$sim(d,q) = \frac{\sum_{j=1}^{n}\sum_{x\in Q} h(x,D_j)}{\sum_{j=1}^{n}\sum_{x\in Q} h(x,Q_j)} \tag{2}$$

The function $h(x, D_j)$, in the equation (2), returns a weight for the j-gram x with respect to the set of j-grams (D_j) in the passage and is defined by:

$$h(x,D_j) = \begin{cases} \sum_{k=1}^{|x|} W_x & if\ x \in D_j \\ 0 & otherwise \end{cases} \tag{3}$$

A more detailed description of the system JIRS can be found in [3].

4 Approach used

The objective was to use the JIRS PR system to detect plagiarism between a candidate patent and any other invention described in the prior art. We suppose that a high similarity value between the candidate patent and another patent in the collection corresponds to the fact that the candidate patent does not represent an original invention. A problem in carrying out this comparison is that JIRS was designed for the QA task, where the input is a question: the JIRS model was not developed to compare a full document to another one but only a sentence (the question) to documents (the passages). Therefore, it was necessary to determine a strategy to summarize the candidate patent in a sequence of words that could be used as a query for JIRS. The summarization technique is based on the random walks method proposed by Hassan et al. [4]: the query is composed by the title of the patent followed by the most relevant n-grams composed by the heaviest terms, according to the weights assigned using the random walks method, assuming a window size of 2 words.

For instance, consider the patent EP-1445166 “*Foldable baby carriage*”, having the following abstract:

> “A folding baby carriage (20) comprises a pair of seating surface supporting side bars (25) extending back and forth along both sides of a seating surface in order to support the seating surface from beneath. Each seating surface supporting side bar (25) has a rigid inward extending portion (25a) extending toward the inside so as to support the seating surface from beneath, at a rear portion thereof. The inward extending portion (25a) is formed by bending a rear end portion of the seating surface supporting side bar (25) toward the inside.”

The random walks method extracts the relevant n-gram *seating surface* from the patent document, composed by the heaviest terms occurring in the document. The resulting query is “Foldable baby carriage, seating surface”.

Another problem was to transform the patents in documents that could be indexed by JIRS. In order to do so, we decided to eliminate all the irrelevant information, extracting from each document its title and the description in the original language in which it was submitted. Each patent has also an identification number, but often the identification number is used to indicate that the present document is a revision of a previously submitted document: in this case it is necessary to examine all documents that are part of a same patent and remove them from the collection. With these transformations we obtained a database that was indexed by the search engine JIRS, in which each of the patents was represented by a single passage.

5 Results

We submitted one run for the task size S (500 topics), obtaining the following results:

Table 1. Result for the submitted run. P: Precision, R:Recall.

P	R	MAP	nDCG
0,0016	0,2547	0,0289	0,3377

6 Conclusions

The obtained results were not satisfactory, possibly due to the reduction process carried out on the provided corpus; however we believe that the assumptions made in the approximation still constitute a valid approach, capable of returning appropriate results; in the future, we will attempt to study how to reduce the database size in order to delete as less amount of relevant information as possible.

The development of the queries regarding each of the patents is one of the weaknesses which must be taken into account for future participations: it will be necessary to refine or improve the summarization process and to compare this model to other summarization models and other standard similarity measures between documents.

Acknowledgements

The work of the first author has been possible thanks to a scholarship funded by Maat Knowledge in the context of the joint project with the Universidad Politécnica de Valencia “Módulo de servicios semánticos de la plataforma G”. We also thank the TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 research project.

References

[1] Callan, J. P. 1994. Passage-level evidence in document retrieval. In Proceedings of the 17th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Dublin, Ireland, July 03 - 06, 1994). W. B. Croft and C. J. van Rijsbergen, Eds. Annual ACM Conference on Research and Development in Information Retrieval. Springer-Verlag New York, New York, NY, 302-310.

[2] Buscaldi D., Sanchis, E., Rosso, P. N-gram vs. Keyword-based Passage Retrieval for Question Answering. Lecture Notes in Computer Science. Vol. 4730 pp.377-384 , 2007

[3] Buscaldi D., P. Rosso, JM Gomez, E. Sanchis Answering Questions with an n-gram based Passage Retrieval Engine. Journal of Intelligent Information Systems (82), 2009.

[4] Hassan S., Mihalcea R., Banea C., Random-Walk TermWeighting for Improved Text Classification. Department of Computer Science University of North Texas.