NLEL-MAAT at CLEF-IP

Santiago Correa and Davide Buscaldi and 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 *CLEF-IP 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.

1 Introduction

The *CLEF-IP 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 [3]. 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 *CLEF-IP 2009* data.

JIRS[1] is an open source *PR* system which was developed at the *Universidad Politécnica de Valencia* (*UPV*), primarily for *Question Answering* (*QA*) tasks. 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 *Question Answering* tracks within the *CLEF Campaign*, *JIRS* proved to be superior in *PR* performance [1] to the *Lucene*[2] open source system. In the following sections, we explain the main concepts of *JIRS* system and show how we adapted JIRS in order to tackle the CLEF-IP retrieval task; in Section 5 we discuss the obtained results; and finally in Section 6 we draw some conclusions.

2 Intellectual Property Task

The main task of the CLEF-IP track consists in finding the prior art for a given patent. The corpus is composed by documents from the *European Patent*

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

Organization (*EPO*)[3] published between 1985 and 2000, a total of 1,958,955 patent documents relating to 1,022,388 patents. The provided documents are encoded in *XML* format, emphasizing these sections: title, language, summary and description, in which our approach can work properly. This supposes the omission of several fields of interest, like IPC class field, and thus a significant loss of information. 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 1,000 documents with their score ranking.

3 The passage retrieval engine JIRS

The *passage retrieval* system *JIRS* is based on n-grams (an n-gram is a sequence of n adjacent words). *JIRS* has the ability to find word sequences structures 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 word sequences in the collection and it gives them a weight in relation to the amount and weights of the n-grams that appear in the query. For instance, consider the two next pasages: *"... braking system consists of disk brakes..."* and *"... anti-lock braking system developed by..."*. If you use a standard search engine, like *Yahoo* or *Lucene*, to search articles related to the phrase "anti-lock braking system", the first passage would obtain a higher weight due to the occurrences of the words containig 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 bigger n-gram, according to the corresponding sub n-gram presents in the query, and assign to it a weight equal to the sum of all term weights. The weight of each term is set to [1]:

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

Where n_k is the number of passages in which the term evaluated appears and N is the total number of passages in the system and k varies between 1 and the number of words in the vocabulary of the corpus.

Once a method for assigning a weight to n-grams has been defined, the next step is to define a measure of similarity between passages. The measure of similarity between a passage (d) and a query (q) is defind as follow:

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

Where Q is the set of n-grams of the passage that are in the query and do not have common terms with any other n-gram. 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:

[3] http://www.epo.org/

$$h(x, D_j) = \begin{cases} \sum_{k=1}^{j} w_k & \text{if } x \in D_j \\ 0 & otherwise \end{cases} \quad (3)$$

A more detailed description of the system *JIRS* can be found in [2].

4 Adaptation of JIRS to the Task

The objective was to use the *JIRS PR* system to detect plagiarism of ideas between a candidate patent and any other invention described in the prior art. We hypothesized 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 query: the *JIRS* model was not developed to compare a full document to another one but only a sentence (the query , preferably short, in which terms are relevant to user needs) to documents (the passages). Therefore it was necessary to summarize the abstract of the candidate patent in order to pass it to *JIRS* as query (i.e., a sequence of words). 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 set of words extracted from the patent abstract using this method.

Consider for instance 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 abstract. The resulting query is "Foldable baby carriage, seating surface".

Another problem was to transform the patents into documents that could be indexed by *JIRS*. In order to do so, we decided to eliminate all the irrelevant information to the purpose of passage similarity analysis, 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 we 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. Due to the corpus is provided in three languages, we decided to implement three search systems, one for each language. Therefore, the input query for each system is given in the language the system was developed. To translate all queries we used the *Google Translation Tool*[4]. For each query we obtained a list of relevant patents by each of the 3 search engines. Finally, we selected the 1,000 better ranked patents. The architecture of our multilingual approach is illustrated in Figure 1.

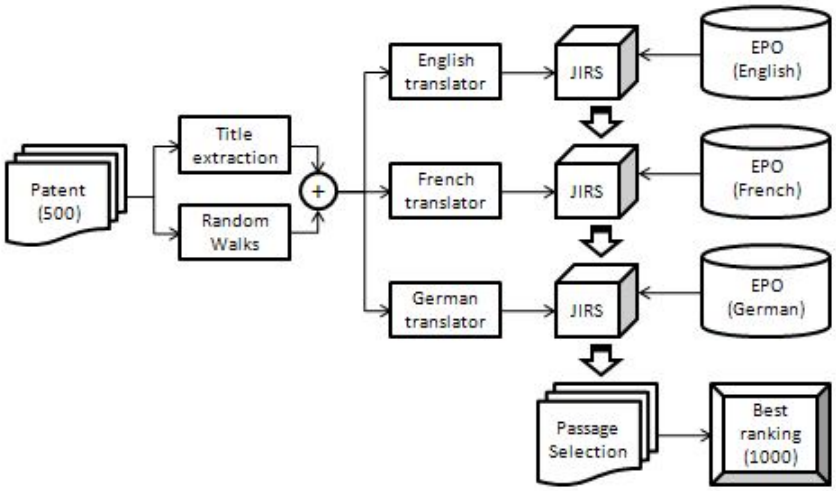

Fig. 1. Architecture of NLEL-MAAT multilingual system

5 Results

We submitted one run for the task size S (500 topics), obtaining the following results [5]:

Table 1. Results obtained for the IP competition by the *JIRS*-based system; P: Precision, R: Recall, nDCG: Normalized Discounted Cumulative Gain, MAP: Mean Average Precision

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

[4] http://www.google.com/language_tools?hl=en

In general, most of the results obtained by the participants were low, due to the complexity of the *IP* task. We have to emphasize that with an approach as simple as the one we have proposed, we have obtained results were not too far from the ones obtained by the best systems, the best system achieve a Precision@100 measure of 0.0317 while our system achieve a Precision@100 measure of 0.0076. From a practical viewpoint, our aim was to apply the simple *JIRS*-based system in order to filter out non-relevant information with respect to the prior art of a patent. This allows to sensibly reduce the size of the data set to investigate eventually employing a more formal approach.

6 Conclusions

The obtained results were not satisfactory, possibly due to the reduction process carried out on the provided corpus, this allows us to be efficient in terms of performance but involves the loss of important information such as the IPC class, inventors, etc.; 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 little 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 as well as similarity measures that include other parameters than the n-grams.

The PLN's approach to the development of the experiments is affected by the necessity to use translation tools which degenerate the quality of the information that is analyzed. We think that an approach from the point of view of classification can throw better results.

Acknowledgments The work of the first author has been possible thanks to a scholarship funded by Maat Gknowledge in the context of the 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. Davide Buscaldi, José Manuel Gómez, Paolo Rosso, and Emilio Sanchis. N-Gram vs. Keyword-Based Passage Retrieval for Question Answering. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *CLEF*, volume 4730 of *Lecture Notes in Computer Science*, pages 377–384. Springer, 2006.

2. Davide Buscaldi, Paolo Rosso, José Manuel Gómez, and Emilio Sanchis. Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, (82):Online First, 2009. ISSN: 0925-9902 (Print) 1573-7675 (Online). DOI: 10.1007/s10844-009-0082-y.
3. James P. Callan. Passage-level evidence in document retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310, New York, NY, USA, 1994. Springer-Verlag New York, Inc.
4. Samer Hassan, Rada Mihalcea, and Carmen Banea. Random-Walk Term Weighting for Improved Text Classification. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 242–249, Washington, DC, USA, 2007. IEEE Computer Society.
5. Giovanna Roda, John Tait, Florina Piroi, and Veronika Zenz. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.