Passage retrieval and intellectual property in legal texts

Santiago Correa, Davide Buscaldi, Paolo Rosso
Natural Language Engineering Lab., EliRF Research Group
Dept. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain
{scorrea, dbuscaldi, prosso}@dsic.upv.es, http://users.dsic.upv.es/grupos/nle

Alfonso Rios
Maat Knowledge, Spain
arios@mat-g.com, http://www.maat-g.com/

Question Answering (QA) can be viewed as a particular form of Information Retrieval (IR), in which the amount of information to return is the minimum required to satisfy the user needs expressed by a specific question such as: "Where is the Europol Drugs Unit?"[1]. A Passage Retrieval (PR) system is an IR system which, given a list of keywords (e.g.: "Electricity," "Motor", etc..) or a question such as the previous one, returns fragments of texts (passages) that are relevant to the user needs.

The *Cross-Language Evaluation Forum*[2] (CLEF), organises competitions for the assessment of multilingual information retrieval systems. In CLEF-2009 edition, due to the growing interest in natural language processing of legal texts from both the university and the business sector, tracks such as ResPubliQA[3] and IP[4] have been organised. We have participated in both tracks in the framework of the collaboration between the Natural Language Engineering Lab. of the Universidad Politécnica de Valencia (UPV) and the *Maat Knowledge* enterprise. In order to address both tracks on QA in legal texts and on Intellectual Property (IP) of patent retrieval, we have used the JIRS (JAVA Information Retrieval System) search engine, a freely available[5] PR system which has been developed at the UPV [1]. The results have been sent to the tracks organisers and will be presented at CLEF-2009 along with the ones of the other teams that have participated in the two tracks.

In the following sections we describe the main concepts of the JIRS system and how it has been applied to the ResPubliQA and IP tracks of CLEF-2009.

1. Passage retrieval system JIRS

Most of nowadays passage retrieval systems are not oriented to the specific question answering problem, because they only take into account the keywords of the question in order

[1] Question from ResPubliQA@CLEF-2009
[2] www.clef-campaign.org/
[3] http://celct.isti.cnr.it/ResPubliQA/
[4] http://www.ir-facility.org/the_irf/clef-ip09-track
[5] http://sourceforge.net/projects/jirs/

to obtain the relevant passages. JIRS is a PR system based on n-grams (an n-gram is a sequence of n adjacent words extracted from a sentence or a question.) JIRS is based on the premise that in a large collection of documents, an n-gram associated with a question must be found in this collection at least once. The PR system has the ability to find structures of questions in a large collection of documents quickly and efficiently through the use of different n-grams models. JIRS searches for all possible n-grams of the question in the collection and it quantifies them in relation to the n-grams quantity and weight that appear in these passages. For example, let us suppose that we have a database of publications of a newspaper. Using the JIRS system we aim at finding in the document of the collection an answer to a question such as “Who is the president of Colombia?”. For instance, the system could retrieve the following two passages: "... Álvaro Uribe is the president of Colombia ...” and “...Giorgio Napolitano is the president of Italy...”. Of course, the first passage should be given more importance because it contains the 5-gram “is the president of Colombia”, whereas the second passage contains only the 4-gram “is the president of”. In order to calculate the n-gram weight of each passage, first of all we need to identify the most relevant n-gram and assign to it a weight equal to the sum of the weights of all its terms. The weight of each term is set to:

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

Where n_k is the number of passages in which the term appears and N is the total number of passages. A more detailed description of the system JIRS can be found in [1].

2. Passage retrieval for question answering

ResPubliQA@CLEF-2009 competition address the problem of question answering in legal texts. Given a pool of 500 independent natural language questions, each system must return the passage (not the exact answer) which answers each question from the JRC-Acquis[6] collection of EU documentation where both questions and documents are translated and aligned for a subset of languages.

In order to use the JIRS system in this QA track, we had to analyse and transform the documents of the collection for indexing them in the JIRS search engine. The collection of the competition is made of documents in XML format, each one divided into paragraphs delimited by the tag <p>. Therefore, each paragraph has been defined as a document, tagged with the name of the document where it is contained and the paragraph number that corresponds to it. Once all the documents have been extracted from the collection, they have been indexed in JIRS according to the language that has been analysed. Once obtained the database indexed by JIRS, we had searched for the answer to each question of the track (see example in the previous page). For each question, the system has returned a list with the most likely documents where an answer to the question was found, according to the way JIRS works. In an additional experiment, we made use of the parallel collection provided for the competition by obtaining a list of answers in different languages (Spanish, English, Italian and French), choosing as best answer the one better ranked by JIRS and then translating all of them to a single language. A detailed description of how the system JIRS has been used in this QA track can be found in [2].

[6] http://langtech.jrc.it/JRC-Acquis.html

3. Passages retrieval for intellectual property

The CLEF IP track is coordinated by *Information Retrieval Facility*[7] (IRF) and *Matrixware*[8]. Its aim is to investigate IR techniques for patent retrieval in order to search for the prior art of a patent on a certain topic in order to determine whether or not a certain degree of plagiarism of ideas occurred. The track provided a collection of more than 1M patent documents, mainly derived from European Patent Office sources, in three languages: English French and German. Queries and relevance judgements have been produced manually by Intellectual Property experts (using a set of queries given by themselves) and automatically, using patent citations from seed patents.

The set of 500 patents in *xml* format contained information that was not useful. Therefore, the first step has been to eliminate this type of information. In addition, patents have a identification number which makes them unique, although it is possible to find different versions of a unique patent. Therefore, we had to eliminate this kind of repeated information. Last, we had to remove stop words from the documents. At the end of this pre-process, we obtained a smaller size collection, which could be indexed by the JIRS search engine. To ask JIRS for the related patents, we had to build the related words sequence to each patent, considering from each of the 500 patents its title and its relevant terms obtained after a summarization technique [3]. For each patent, the information of the title and its relevant terms was concatenated and given to JIRS as a words sequence.

A detailed description of how the system JIRS has been used in the task can be found in [4].

Acknowledgement

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

References

1. Buscaldi D., Rosso P., Gómez J.M., Sanchis E. Answering Questions with an n-gram based Passage Retrieval Engine. Journal of Intelligent Information Systems (82), 2009 (in press).
2. Correa S., Buscaldi D., Rosso P. NLEL-MAAT at CLEF-ResPubliQA. Proc. 10th Int. Cross-Language Evaluation Forum CLEF-2009 working notes (to be published in September 2009).
3. Hassan S., Mihalcea R., Banea C., Random-Walk TermWeighting for Improved Text Classification. IEEE International Conference on Semantic Computing, ICSC-2007, 2007.
4. Correa S., Buscaldi D., Rosso P. NLEL-MAAT at CLEF-IP at CLEF-ResPubliQA. Proc. 10th Int. Cross-Language Evaluation Forum, CLEF-2009 working notes (to be published in September 2009)

[7] http://www.ir-facility.org/the_irf/

[8] http://www.matrixware.com/