NLEL-MAAT at ResPubliQA

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 *QA@CLEF-2009* competition. We used the *JIRS* passage retrieval system, which is based on redundancy, with the assumption that it is possible to find the response to a question in a large enough document collection. The retrieved passages are ranked depending on the number, length and position of the question n-gram structures found in the passages. The best results were obtained in monolingual English, while the worst results were obtained for French. We suppose the difference is due to the question style that varies considerably from one language to another.

1 Introduction

An open-domain *Question Answering* (*QA*) system can be viewed as a specific *Information Retrieval* (*IR*) system, in which the amount of information retrieved is the minimum amount of information required to satisfy a user information need expressed as a specific question, e.g.: "Where is the Europol Drugs Unit?". Many *QA* systems are based on *Passage Retrieval* (*PR*) [6, 4]. A *PR* system is an *IR* system that returns parts of documents (passages) instead of complete documents. Their utility in the *QA* task is based on the fact that in many cases the information needed to answer a question is usually contained in a small portion of the text [3].

In the 2009 edition of *CLEF*, the competition *ResPubliQA*[1] has been organized, a narrow domain *QA* task, centered on the legal domain, given that the data is constituted by the body of *European Union* (*EU*) law. Our participation in this competition has been based on the *JIRS*[2] open source *PR* system, which has proved to be able to obtain better results than classical *IR* search engines in the previous open-domain *CLEF QA* tasks [1]. In this way we desired to evaluate the effectiveness of this *PR* system in this specific domain and to check our hypothesis that most answers usually are formulated similarly to questions, in the sense that they contain mostly the same sequences of words. In the next section, we describe the characteristics of the task; furthermore, Sect. 3 and 4

[1] For more information about the competition ResPubliQA@CLEF-2009, refer to page: http://celct.isti.cnr.it/ResPubliQA/

[2] http://sourceforge.net/projects/jirs/

explain the main concepts of *JIRS* (*Java Information Retrieval System*) system and we discuss how it has been applied in solving the problem; in Sect. 5 we present the results and finally in Sect. 6 we draw some conclusions.

2 Multiple Language Question Answering Task

In this task, the system receives as input natural language questions about European law, and the system should return a paragraph containing the response from the document collection. This constitutes an important difference with respect to previous *QA* tasks where an exact answer had to be extracted or generated by the system. For this reason we employed just the *JIRS* system instead of the complete *QUASAR QA* system we developed for previous QA@CLEF participations [2].

The document collection is a subset of the *JRC-Acquis corpus* [3] , containing the complete *EU* legislation, including texts between the years 1950 to 2006 (in total 10,700 documents); these documents have been aligned in parallel and were made available to the participants in the following languages: Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish. The corpus is encoded in *XML* format according to the *TEI guidelines*[4] . Each document has a title and is subdivided into paragraphs, each one marked with the "<p>" tag. The test set is composed of 500 questions that must be analyzed by the systems to return a paragraph that contains the answer to the formulated question.

3 The Passage Retrieval Engine JIRS

Many passage retrieval systems are not targeted to the specific problem of finding answers, due to the fact that they only take into account the keywords of the question to find the relevant passages. The information retrieval system *JIRS* is based on n-grams (an n-gram is a sequence of n adjacent words extracted from a sentence or a question) instead of keywords. *JIRS* is based on the premise that in a large collection of documents, an n-gram associated with a question must be found in the collection at least once.

JIRS starts searching the candidate passage with a standard keyword search that retrieves an initial set of passages. These passages are ranked later depending on the number, position and length of the question n-grams that are found in the passages. For example: suppose you have a newspaper archive, using the JIRS system and based on these documents you will find the answer to the question: "Who is the president of Colombia?". 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

[3] http://wt.jrc.it/lt/Acquis/
[4] http://www.tei-c.org/Guidelines/

should have more relevance as it contains the 5-gram "is the president of Colombia", while the second passage contains only the 4-gram "is the president of". 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 its terms. 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 similarity between a passage d and a question q is determined by:

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

Where $h(x, D_j)$ returns a weight for the j-gram x with respect to the set of j-grams (D_j) in the passage:

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

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

4 Adaptation of JIRS to the Task

The data had to be preprocessed, due to the format of the collection employed in *ResPubliQA* competition, a subset of the *JRC-ACQUIS* Multilingual Parallel corpus, this corpus containing the total body of *European Union* (*EU*) documents, of mostly legal nature. In particular, the subset is constituted by documents of 9 out of 22 languages. It consists of approximately 10,700 parallel and aligned documents per language. The documents cover various subject domains: law, politics, economy, health, information technology, agriculture, food and more.

To be able to use the *JIRS* system in this task, the documents were analyzed and transformed for proper indexing. Since *JIRS* uses passages as basic indexing unit, it was necessary to extract passages from the documents. We consider any paragraph included between <p> tags as a passage. Therefore, each paragraph was labeled with the name of the containing document and its paragraph number.

Once the collection was indexed by *JIRS*, the system was ready to proceed with the search for the answers to the test questions. For each question, the system returned a list with the passages that most likely contained the answer to the question, according to the *JIRS* weighting scheme. The architecture of the monolingual *JIRS*-based system is illustrated in Fig. 1. In an additional experiment, we used the parallel collection to obtain a list of answers in different languages (Spanish, English, Italian and French). The idea of this approach is

based on the implementation of 4 monolingual *JIRS*-based systems, one for each language, which will have as input the set of questions in the respective language. For this purpose we used a tool (Google Translator[5]) to translate the entire set of questions into the same language. Later choosing as the best answer the one that obtained the best score by the system and subsequently taking the identifier of each paragraph (answer) for retrieving the aligned paragraph in the target language. The architecture of the multilingual *JIRS*-based system is illustrated in Fig. 2.

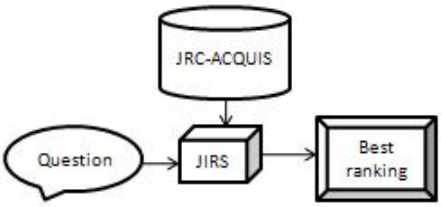

Fig. 1. Architecture of NLEL-MAAT monolingual system

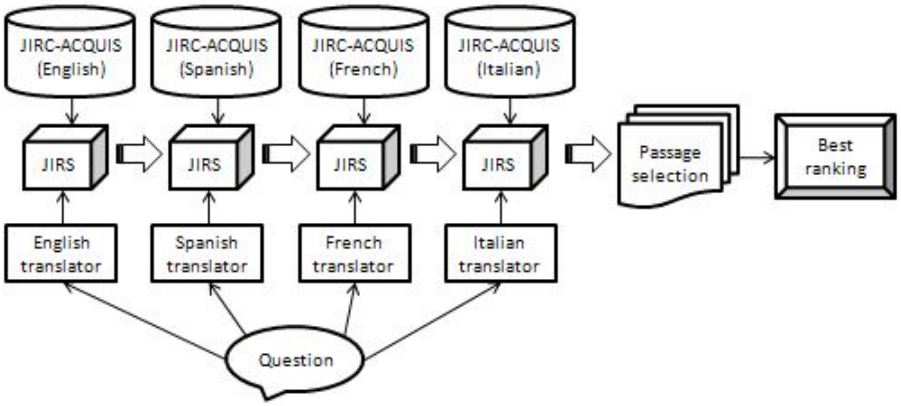

Fig. 2. Architecture of NLEL-MAAT multilingual system

5 Results

We submitted four "pure" monolingual runs for the following languages: English, French, Italian and Spanish, and in an additional experiment we exploited the

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

parallel corpus to produce a monolingual Spanish run. This experiment consisted in searching the question in all languages, and selecting the passage with the highest similarity; finally, the returned passage was the Spanish alignment of this best passage. In Table 1 we show the official results for the submitted runs [5].

Table 1. Results for submitted runs to ResPubliQA, Ans.: Answered, Unans.: Unanswered, A.R.: Answered Right, A.W.: Answered Wrong, U.R.: Unanswered Right, U.W.: Unanswered Wrong, U.E.: Unanswered Empty, Accuracy: Accuracy measure, c@1

Task	Ans.	Unans.	A.R.	A.W.	U.R.	U.W.	U.E.	Accuracy	c@1
en-en	498	2	287	211	0	0	2	0.57	0.58
fr-fr	489	11	173	316	0	0	11	0.35	0.35
es-es	495	5	173	322	0	0	5	0.35	0.35
it-it	493	7	256	237	0	5	2	0.51	0.52
es-es2	466	34	218	248	0	0	34	0.44	0.47

From Fig. 3 we can see that the result obtained in English were particularly good, while in French and Spanish the percentage of wrong answers is very high. We did not expect this behavior for Spanish, since *JIRS* was developed specifically for the Spanish *QA* task. Maybe the poor behavior was due to the peculiarity of the *JRC Acquis* corpus, containing, for instance, many anaphoras. On the other hand, we expected the French to be the language in which the system obtained the worst results because of the results of the system at previous QA competitions. The Spanish-multilingual approach allowed to reduce the wrong answers by 23% (from 322 to 248) and increase the number of right ones by 26% (from 173 to 218).

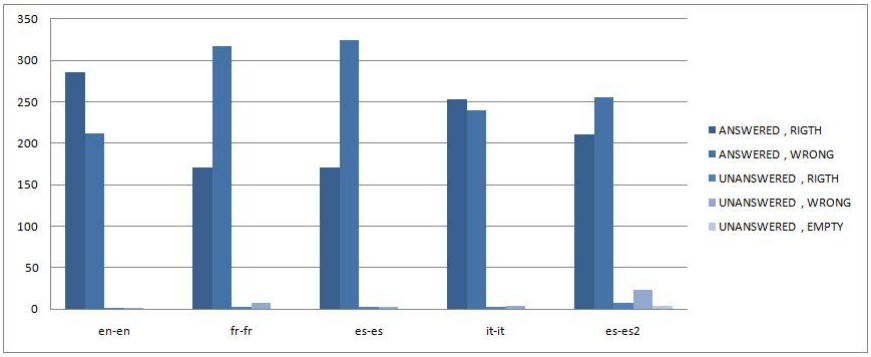

Fig. 3. Comparative graph for all the sumitted runs

6 Conclusions

The difference between the best results (for English) and the worst ones (for French) is 22 percent points in accuracy. This may reflect the different way of formulating questions in each language. In a comparison with other teams' results, we obtained excellent results, proof of this is the best rating in three of the four tasks we participated in. The only task in which the NLEL-MAAT system was not ranked first is the monolingual task en-en. However, the system ranked second with just a difference of 0.03 in the c@1 measure and 0.04 in the *Accuracy* measure. Moreover it is also important to note that due to the language independence of *JIRS* we have participated and obtained very good results in all the tasks. Due to the improvement obtained using the parallel data set (es-es2 task) with respect to the Spanish monolingual task (es-es), we plan to employ this approach also for the other languages.

Acknowledgments The work of the first author was made possible by 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.
4. Günter Neumann and Bogdan Sacaleanu. Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answering System. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *CLEF*, volume 3491 of *Lecture Notes in Computer Science*, pages 411–422. Springer, 2004.
5. Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forascu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, and Petya Osenova. Overview of respubliqa 2009: Question answering evaluation over european legislation. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

6. Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Cnference on Research and Development in Information Retrieval*, pages 41–47, New York, NY, USA, 2003. ACM.