

NLEL-MAAT at CLEF-ResPubliQA

Santiago Correa, Davide Buscaldi, Paolo Rosso.
NLE Lab, ELiRF Research Group, DSIC,
Universidad Politécnica de Valencia, Spain.
{scorrea, dbuscaldi, proso}@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-grams 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.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; I.2 [Artificial Intelligence]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Question Answering, Information Retrieval and Passage Retrieval.

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) [1,2]. 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 reply a question is usually contained in a small portion of the text [3].

In the 2009 edition of CLEF, the competition ResPubliQA¹ has been organized, consisting in 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 JIRS² 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 [4]. In this way we desired to evaluate the effectiveness of this PR system in this specific domain and to check our hypothesis that answers usually are formulated similarly to questions. In the next section we describe the characteristics of competition; furthermore, paragraphs 3 and 4 explain the main concepts of JIRS system and we discuss how it has been applied in solving the problem, continuing, in paragraph 5, with the results and the conclusions.

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/>

2 Multiple Language Question Answering Task

In this task the systems receive in input natural language questions about knowledge of European law, and these, in turn, should return a paragraph containing the response from the documents collection. This constitutes an important difference with respect to previous QA tasks where the answer had to be extracted by the system. For this reason we employed just the JIRS system instead of the complete QUASAR QA system we developed for previous participations.

The document collection is composed by 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 competition 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 guidelines⁴. Each document have a title and is subdivided in a series of phrases, 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

Nowadays, 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 than 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 this 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. E.g.: suppose you have a publications database of a newspaper, 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 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-grams weight of each passage, first of all you must identify the most relevant n-gram and assign to it an equal weight to the all weights sum of the terms. The weight of each term is set to:

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

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:

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

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:

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

A more detailed description of the system JIRS can be found in [5].

³ <http://wt.jrc.it/lt/Acquis/>

⁴ <http://www.tei-c.org/Guidelines/>

4 Adaptation of JIRS to the Task

The data had to be preprocessed, due to the format of the collection employed in the ResPubliQA@CLEF-2009 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 of 10,700 parallel and aligned documents per language. The documents cover various subject domains: economy, health, information technology, law, agriculture, food, politics and more.

To be able to use the JIRS system in this task, the documents were analysed 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 $\langle p \rangle$ tags as a passage. Therefore, each paragraph was labelled 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 questions provided by the competition. 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. In an additional experiment, we used the parallel collection to obtain a list of answers in different languages (Spanish, English, Italian and French), choosing as the best answer the one that obtained the best score by JIRS and subsequently taking the identifier of each paragraph (answer) for retrieving the aligned paragraph in the target language.

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 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.

Table 1. Results for submitted runs. Ans.: Answered, Unans.: Unanswered, A.R.: Answered Righ, A.W.: Answered Wrong, U.R.: Unanswered Righ, U.W.: Unanswered Wrong, U.E.: Unanswered Empty, Overall: Overall accuracy, PACD: Proportion of answers correctly discarded, c@1: c@1 measure.

Task	Ans.	Unans.	A.R.	A.W.	U. R.	U. W.	U. E.	Overall	PACD	c@1
en-en	498	2	286	212	1	1	0	0,57	0,5	0,57
fr-fr	488	11	171	317	3	8	0	0,35	0,73	0,35
es-es	495	5	171	324	2	3	0	0,35	0,6	0,35
it-it	493	7	253	240	3	4	0	0,51	0,57	0,51
es-es2	466	34	211	255	7	23	4	0,44	0,68	0,45

From Figure 1 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 behaviour for the Spanish language, since JIRS was developed specifically for the Spanish QA task. 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.

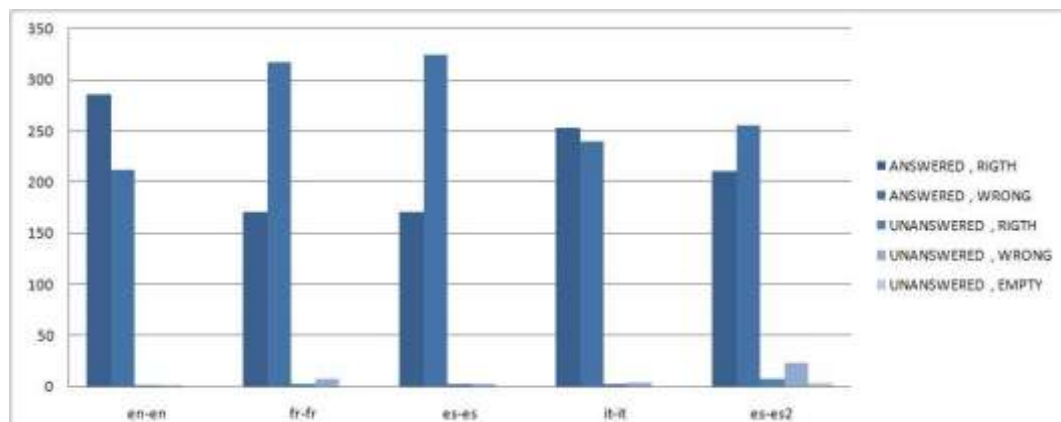


Fig. 1. Comparative graph for all the submitted runs.

6 Conclusions

The difference between the best results (for English) and the worst ones (in French) is of 22%. This may reflect the different way of formulating questions in each language. The abandonment of the Question Classification and Answer Extraction phases, with respect to our previous participation in the CLEF QA tasks, did not result particularly useful. In many cases, the right answer was available in a passage ranked lower, and this could have been detected by checking the question type and determining if the passage contained a candidate answer. We plan to re-introduce these modules for further participations, at least for the validation of the returned passage, that is, to check if it contains the answer or not.

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 Politécnic 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] Tellex, S., Katz, B., Lin, J., Fernandes, A., and Marton, G. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in Informaion Retrieval (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM, New York, NY, 41-47. DOI= <http://doi.acm.org/10.1145/860435.860445>
- [2] Neumann, G., & Sacaleanu, B. (2005). Experiments on robust NL question interpretation and multi-layered document annotation for a cross-language question/answering system. In Multilingual information access for text, speech and images. LNCS 3491, pp. 411-422
- [3] 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.
- [4] Buscaldi D., P. Rosso, JM Gomez, E. Sanchis: N-Gram vs. Keyword-Based Passage Retrieval for Question Answering. CLEF 2006: 377-384
- [5] 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.