Re-ranking of Yahoo Snippets with the JIRS Passage Retrieval System

José Manuel Gómez, Paolo Rosso, Emilio Sanchis
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Valencia, Spain
{jogomez,prosso,esanchis}@dsic.upv.es

Abstract

Passage Retrieval (PR) systems are used as first step of the actual Question Answering (QA) systems. Usually, PR systems are traditional information retrieval systems which are not oriented to the specific problem of QA. In fact, these systems only search for the question keywords. JIRS Distance Density n-gram system is a QA-oriented PR system which has given good results in QA tasks when this is applied over static document collections. JIRS is able to search for the question structure in the document collection in order to find the passages with the greatest probability to contain the answer. JIRS is a language-independent PR system which has been already adapted to a few non-agglutinative European languages (such as Spanish, Italian, English and French) as well as to the Arabic language. A first attempt to adapt it to the Urdu Indian language was also made. In this paper, we investigate the possibility of basing on the web the JIRS retrieval of passages. The experiments we carried out show that JIRS allow to improve the coverage of the correct answers re-ranking the snippets obtained with Yahoo search engine.

1 Introduction

A QA system is an application that allows to a user to make questions in natural language in order to look for the correct answer in a non-structured document collection. In the multilingual QA tasks, it is very important to use methodologies of document (or passage) retrieval as independent of the language as possible.

Document or passage retrieval is typically used as the first step in current QA systems [Corrada-Emmanuel *et al.*, 2003]. In most of the QA systems, classical PR systems are used [Magnini *et al.*, 2001; Aunimo *et al.*, 2004; Vicedo *et al.*, 2003; Neumann and Sacaleanu, 2004]. The main problem of these QA systems is that they use PR systems which are adaptations of classical document retrieval systems instead of being oriented to the specific problem of QA. These systems use the question keywords to find relevant passages. Other PR approaches are based on Natural Language Processing (NLP) [Ahn *et al.*, 2004; Greenwood, 2004; Hess, 1996; Liu and Croft, 2002]. These approaches have the disadvantage to be very difficult for adaptation to other languages or to multilingual tasks.

The strategy of Castillo, Brill and Buchholz [Del-Castillo-Escobedo *et al.*, 2004; Brill *et al.*, 2001; Buchholz, 2001] is to search the obviousness of the answer in the Web. They run the user question into a Web search engine (usually Google[1]) with the expectation to get a passage containing the same expression of the question or a similar one. They suppose that due to the high redundancy[2] of the Web, the answer will be written in different ways but including the complete question expression. Unfortunately, the matter is that very often the answer does not appear in a context similar to the question expression. To increase the possibility to find relevant passages they make reformulations of the question, i.e., they move or delete terms to search other structures with the same question terms. For instance, if we move the verb of the question *Who is the President of India?* and we delete the question term *Who*, we obtain the query *the President of India is*. Thanks to the redundancy, we might find a passage with the structure *the President of India is* ***APJ Abdul Kalam***. Brill makes the reformulations carrying out a Part Of Speech analysis of the question and moving or deleting terms of specific morphosyntactic categories. Castillo makes instead the reformulations doing certain assumptions about the verb position and the prepositional phrases boundaries in the question. The problem of these systems is that all possible reformulations of the question are not taken into account.

With the methods used by Brill and Castillo it would be very costly to realize all possible reformulations since every reformulation must be searched by search engine.

In this paper we describe the JAVA Information Retrieval System[3] (JIRS) adapted to work on the Web. In order to do it, JIRS makes use of Yahoo[4] search engine in the first steps and then it re-ranks the returned snippets using the Distance Density n-gram model. JIRS showed to be able to return the most probable snippets containing the answer.

[1] www.google.com

[2] Certain repetition of the information contained in the collection of documents or Web, which allows, in spite of the loss of a part of the information, to reconstruct its content

[3] http://jirs.dsic.upv.es/

[4] http://www.yahoo.com

The remainder of this work is structured as follows. In Section 2, the general architecture of the system together with the Distance Density n-gram model is described. In Section 3 the metric measures are presented. In Section 4 we discuss the obtained results. Finally, in Section 5 we draw conclusions and we present some the future works.

2 Description of the JIRS PR system

JIRS Distance Density n-gram system [Gómez *et al.*, 2006] makes a systematical search of all question structures in order to find pieces of text with the greatest probability to contain the correct answer. In its web-based version, JIRS uses the Yahoo search engine as first step. Next, it searches all relevant n-grams in the retrieved snippets and then it rates them according to the weight of the n-grams appeared in these snippets.

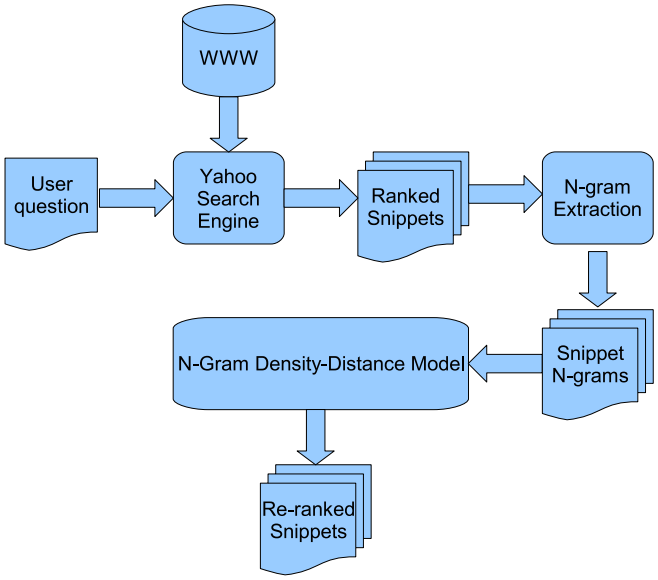


Figure 1: Main structure of JIRS Distance Density n-gram system (web-based version)

In Figure 1 we can observe the main structure of the system. The *Yahoo Search Engine* module performs a search with the user question in order to find in the Web the relevant snippets (i.e., pieces of text) with the question keywords.

With the 1000 most relevant snippets returned by Yahoo, the system extracts the 1-grams, 2-grams and so forth up to the n-gram (where n is the number of question terms). These n-grams are only compounded by question terms. Then, the snippet set of n-grams are compared with the question using a *Distance Density n-gram* model. This model finds question structures in the snippets and gives a higher similarity value to those snippets that contain more grouped structures. This similarity value is calculated by:

$$Sim(p,q) = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{\forall x \in P} h(x) \frac{1}{d(x, x_{max})} \qquad (1)$$

Let Q be the set of n-grams of p composed only by question terms. Therefore, we define $P = \{x_1, x_2, ..., x_M\}$ as a sorted subset of Q that fulfils the following conditions:

1. $\forall x_i \in P$:
   $h(x_i) \geq h(x_{i+1}) \quad i \in \{1, 2, ..., M-1\}$
2. $\forall x, y \in P : x \neq y \Rightarrow T(x) \bigcap T(y) = \emptyset$
3. $\min_{x \in P} h(x) \geq \max_{y \in Q \setminus P} h(y)$

where $T(x)$ is the set of terms of the n-gram x, and $h(x)$ is the function defined by:

$$h(x) = \sum_{k=1}^{j} w_k \qquad (2)$$

where $w_1, w_2, ..., w_{|x|}$ are the term weights of the n-gram x and are calculated by:

$$w_k = 1 - \frac{\log(n_k)}{1 + \log(N)} \qquad (3)$$

where n_k is the number of passages in which the term t_k occurs and N is the number of system passages. As the calculation of the n_k and N values from the Web collection of documents is very difficult, we approximate these values using a static document collection. We have used the Spanish CLEF[5] corpus to obtain these values.

According to the Equation 3, each term has different weight which depends on its relevance. For example, stopwords have the least relevance and the terms that appear only once have the most. These weights give an incentive to those terms which do not appear very often in the document collection. Moreover, the weights should also discriminate the terms against those (e.g. stopwords) that occur often in the document collection.

The $d(x, x_{max})$ is a distance factor between the n-gram x and the n-gram x_{max} and it is calculated by:

$$d(x, x_{max}) = 1 + \ln(1 + L) \qquad (4)$$

where L is the number of terms (including stopwords) between the n-grams. Therefore, the distance factor is equal to 1 when the n-grams appear together and it rises as the distance increases reducing the n-gram weights.

In Figure 2 we can observe an example of this model. The first snippet contains only one question n-gram and its similarity value is the sum of its terms divided by the sum of the weights of all question terms. However, the second snippet has two question n-grams. The greatest n-gram is "*the Croatia*" with a weight of 0.6. The other n-gram is "*capital of*" with a weight of 0.3. We would like to emphasise that the n-grams are formed, only, by question terms and the terms included in greater n-grams do not can included in other n-gram less weighted. That is why the n-gram "*the capital of*" is not taken into account ("*the*" term is already included into the greater n-gram "*the Capital*"). Therefore, the distance between both n-grams is equal to 7. Thefore, the "*capital of*"

[5]Cross Language Evaluation Forum (http://clef.iei.pi.cnr.it/).

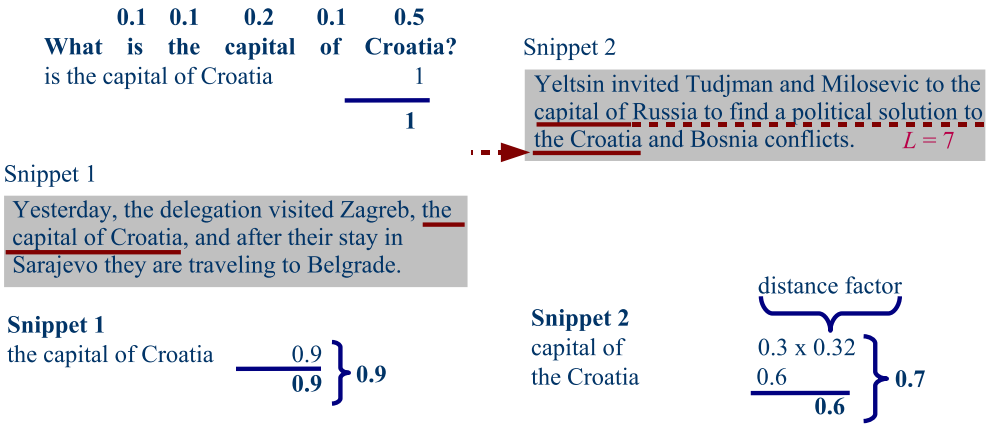


Figure 2: Example of Distance Density n-gram model

weight decreases to 0.1 due to the distance factor. If we calculate the similarity for both snippets, we obtain the value 0.9 for the first snippet and 0.7 for the second one.

In the Distance Density n-gram model, those snippets that contain n-grams with more relevant terms have greater weight than others. Thus, if a n-gram does not contain one of the relevant terms, the weight associated with this n-grams will be diminished much more than the weight of another one which does not include a non-relevant term (e.g. a stopword). Another characteristic of this model is that the similarity value is not affected by the question reformulations. For instance, the n-gram "*is the capital of Croatia*" will have the same weight as "*the capital of Croatia is*" even if does not contain the n-gram for the simple reason that it is formed by the question terms. This aspect is very important for languages whose answer expressions are, normally, reformulations of question terms.

The JIRS system was the core PR system of three QA systems that participated CLEF 2005 and 2006 [Montes *et al.*, 2006; Gómez *et al.*, 2006]. These QA systems obtained the best results in the Spanish and Italian monolingual tasks and in the English-Spanish and Spanish-English multilingual tasks in 2005 and 2006.

3 Evaluation metrics

The experiments detailed in this paper will be evaluated using a metric known as *coverage* (for more details see [Roberts and Gaizauskas, 2004]).

Let Q be the question set, S the all possible snippets which we can obtain from Internet, $A_{S,q}$ the subset of S containing correct answers to $q \in Q$, and $R_{S,q,n}$ be the top n ranked documents in S retrieved by the search engine given a question q.

The $coverage$ of the search engine for a question set Q and the snippet collection S at rank n is defined as:

$$coverage(Q,S,n) \equiv \frac{|\{q \in Q | R_{S,q,n} \cap A_{S,q} \neq \emptyset\}|}{|Q|} \quad (5)$$

The coverage gives the proportion of the question set for which a correct answer can be found within the top n snippets retrieved for each question.

Another metric used in the experiments is the *Mean Reciprocal Rank* (MRR). This measure was defined in [Voorhees, 1999]. Given a set of questions Q, the set of snippets collections S, the subset $A_{S,q}$ of S which contains the correct answers for $q \in Q$, and the set of the first documents $R_{S,q,n}$ of S returned for every question q, so that $R_{S,q,n} = \{s_{q,1}, s_{q,2}, ..., s_{q,n}\}$, the MRR is defined by:

$$mrr(Q,S,n) = \frac{\sum_{\forall q \in Q} rr(q, R_{S,q,n})}{|Q|} \quad (6)$$

where $rr(q, R_{S,q,n})$ is the *Reciprocal Rank* (RR) that depends on the position of the first returned snippet from the result list which contains the answer. Or 0 if the answer is not found in the first n snippets. This function is defined by:

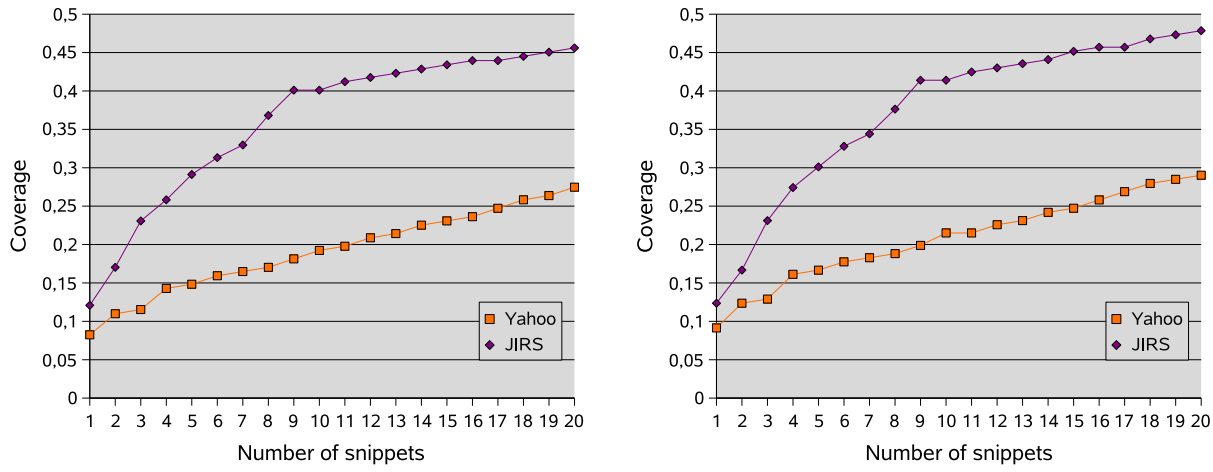
$$rr(q, R_{S,q,n}) = \begin{cases} \frac{1}{i} & \text{si } \exists i | i = \min_{1 \leq j \leq n} j | s_{q,j} \in A_{S,q} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

The *answer redundancy* gives the average number, per question, of snippets within top n retrieved snippets which contain a correct answer. Thefore, the answer redundancy of a retrieval system for a question set Q and the snippet collection S at rank n is defined as:

$$redundancy(Q,S,n) \equiv \frac{\sum_{q \in Q} |R_{S,q,n} \cap A_{S,q}|}{|Q|} \quad (8)$$

4 Preliminary results

We carried out some preliminary experiments on the 200 questions of the CLEF 2005 Spanish QA task. We considered two answer collections developed by two human evaluators using different PR systems in order to obtain a wide range of possible answers for every question. Two different criteria



(a) Strict evaluation

(b) Lenient evaluation

Figure 3: Comparison of coverage: JIRS vs Yahoo

were used to make the two answer collections[6]. The first criterion is a *strict* approach: we have only taken into account the answer given by the CLEF evaluators plus also those answers which we made sure that were correct. The second answer collection was built on the basis of a *lenient* approach and it contains also those answers whose correctness could be arguable because their subjectivity. For instance, for the question "*What is the FARC?*", a strict criterion would be "*Fuerzas Armadas Revolucionarias de Colombia*" but a lenient criterion would "*guerrilla group*" or "*rebellious group*".

In the Figure 3 we can show the improvement of JIRS coverage with respect to Yahoo search engine. In the experiments, the strict and lenient evaluations were used. In both cases, the coverage of JIRS exceeds in a 19% the Yahoo coverage in the first 20 snippets. Moreover, the improve of JIRS is radical even in the first 5 retrieved snippets. Unfortunately, the final coverage, for both systems, is not very high. The experiments we carried out previously over static document collections, achieves instead coverage values of 75% and 90% for the strict and lenient evaluations, respectively [Gómez *et al.*, 2007]. The poor performance obtained is due to the fact that the evaluation answers were obtained from the static CLEF 2005 Spanish corpus (this document collection is composed of documents of the *Agencia EFE* from 1994 to 1995) whereas using the web it was more likely to find snippets containing updated answers. For instance, for the question "*Who is the Primer Minister of Spain?*" the right answer in the Agencia EFE collection is "***José María Aznar***" whereas on the web the actual answer if no specified in 1995, would be "***José Luis Rodríguez Zapatero***". Another possible reason is how snippets are presented by the Yahoo search engine. In fact, them have incomplete sentences and often dots are added before that the answer appears. However, in spite of these inconveniences, we may show that the Distance Density n-gram model of JIRS improves the Yahoo coverage, obtaining an improvement of approximately 20% with respect to the Yahoo search engine.

Table 1 represents the Mean Reciprocal Rank of both systems and evaluations for the first 5 snippets. In the previous figure, we can appreciate that the difference in coverage increases with the number of snippets. JIRS obtains a MRR of 0.07 higher than Yahoo in the first 5 snippets. We believe the difference between the two MRRs could be even greater in case of: (i) solving the incompleteness problem of Yahoo snippets and possibly (ii) building an updated answer collection.

	Strict	**Lenient**
Yahoo	0,105952	0,118459
JIRS	0.179212	0.182796

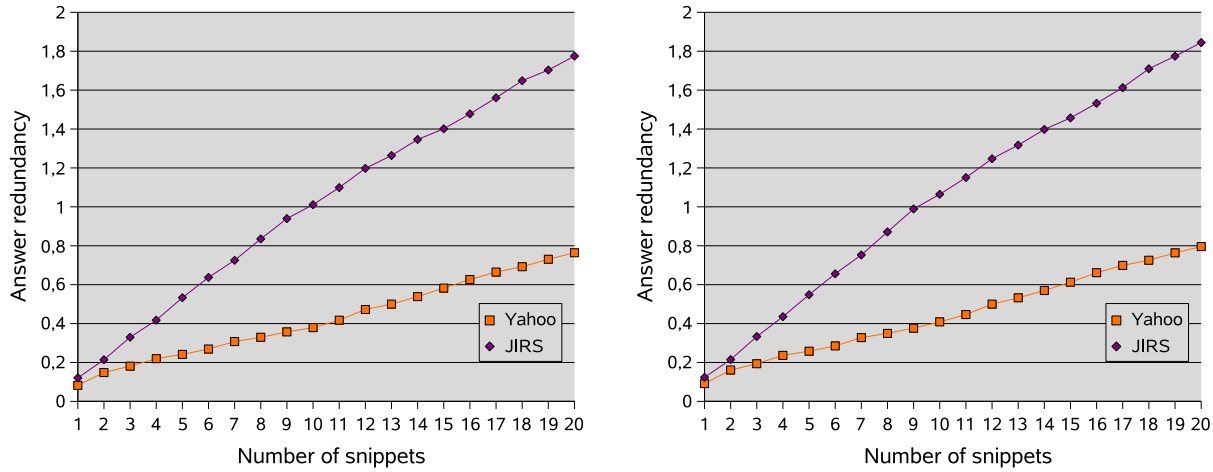
Table 1: MRR for the first 5 snippets

The redundancy of both systems is showed in the Figure 4. We can also note the improvement of JIRS redundancy with respect to Yahoo search engine for both evaluations. JIRS does not have only a better coverage but it has also more redundancy. In fact, JIRS is able to find one answer more for every question than Yahoo.

5 Conclusions and Further Work

In this paper we presented the results of some preliminary experiments we carried out in order to investigate the issue of how JIRS passage retrieved system may to re-rank a search engine snippets in order to make easier the answer extraction. We have seen that the Distance Density n-gram model of JIRS improves both the coverage and the redundancy of the answers.

[6]Both sets of answers can be downloaded from http://jirs.dsic.upv.es.



(a) Strict evaluation

(b) Lenient evaluation

Figure 4: Comparison of answer redundancy between Yahoo and JIRS

Our system has the advantage to be language-independent because it is based on processing the question and the passages without using any knowledge about the lexicon and the syntax of the corresponding language [Gómez *et al.*, 2005]. For this reason, JIRS is very appropriate to find answers in the multilingual document collections and, thefore, in the web. In any non-agglomerative language not many differences between the question and the answer sentences, our system should work very well. At the moment of writing this paper, we are also investigating the possibility of adapting the JIRS PR system to some of the official Indian language. This is the aim of the future two-month visit of the first author of the paper.

As furtherwork we need to overtake the problems we came across in the web-based version of JIRS searching for the answer in the whole document avoid the problem of incomplete snippets with dots.

For the near future, we have decided to resolve the problems which we have found. The first step is to create two answer collections evaluation by human beings from Internet. In this way, we will be able to evaluate JIRS using up-to date answers. The second step is to search question structures using Web pages instead of snippets. Therefore, searching in the whole document we can avoid the problem of the cut sentences in the snippets.

References

[Ahn *et al.*, 2004] Risuh Ahn, Beatrix Alex, Johan Bos, Tiphaine Dalmas, Jochen L. Leidner, and Matthew B. Smillie. Cross-lingual question answering with qed. In *Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004)*, Bath, UK, 2004.

[Aunimo *et al.*, 2004] Lili Aunimo, Reeta Kuuskoski, and Juha Makkonen. Cross-language question answering at the university of helsinki. In *Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004)*, Bath, UK, 2004.

[Brill *et al.*, 2001] Eric Brill, Jimmy Lin, Michele Banko, Susan T. Dumais, and Andrew Y. Ng. Data-intensive question answering. In *The 10th Text REtrieval Conference*, 2001.

[Buchholz, 2001] Sabine Buchholz. Using grammatical relations, answer frequencies and the world wide web for trec question answering. In *The 10th Text REtrieval Conference*, 2001.

[Corrada-Emmanuel *et al.*, 2003] Andrés Corrada-Emmanuel, Bruce Croft, and Vanessa Murdock. Answer passage retrieval for question answering. Technical Report, Center for Intelligent Information Retrieval, 2003.

[Del-Castillo-Escobedo *et al.*, 2004] Alejandro Del-Castillo-Escobedo, Manuel Montes-Gómez, and Luis Villaseñor-Pineda. Qa on the web: a preliminary study for spanish language. In *Proceedings of the Fifth Mexican International Conference in Computer Science (ENC'04)*, Colima, Mexico, 2004.

[Gómez *et al.*, 2005] José Manuel Gómez, Manuel Montes-Gómez, Emilio Sanchis, Luis Villaseñor-Pineda, and Paolo Rosso. Language independent passage retrieval for question answering. In *Fourth Mexican International Conference on Artificial Intelligence MICAI 2005*, Lecture Notes in Computer Science, pages 816–823, Monterrey, Mexico, 2005. Springer Verlag.

[Gómez *et al.*, 2006] José Manuel Gómez, Davide Buscaldi, Empar Bisbal-Asensi, Paolo Rosso, and Emilio Sanchis. *QUASAR: The Question Answering System of the Universidad Politecnica de Valencia*, volume 4022 of *Lecture Notes in Computer Science*, pages 439–448. Springer Verlag, Vienna, Austria, 2006.

[Gómez *et al.*, 2007] José Manuel Gómez, Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. Jirs: Language-independent passage retrieval system: A comparative

study. In *5th International Conference on Natural Language Processing 2006*, Hyderabad, India, 2007.

[Greenwood, 2004] Mark A. Greenwood. Using pertainyms to improve passage retrieval for questions requesting information about a location. In *SIGIR*, 2004.

[Hess, 1996] Michael Hess. The 1996 international conference on tools with artificial intelligence (tai'96). In *SIGIR*, 1996.

[Liu and Croft, 2002] X. Liu and W. Croft. Passage retrieval based on language models. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, 2002.

[Magnini *et al.*, 2001] Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. Multilingual question/answering: the diogene system. In *The 10th Text REtrieval Conference*, 2001.

[Montes *et al.*, 2006] Manuel Montes, Luis Villaseñor, Manuel Pérez, José Manuel Gómez, Emilio Sanchis, and Paolo Rosso. *A Full Data-Driven System for Multiple Language Question Answering*, volume 4022 of *Lecture Notes in Computer Science*, pages 439–448. Springer Verlag, Vienna, Austria, 2006.

[Neumann and Sacaleanu, 2004] Günter Neumann and Bogdan Sacaleanu. Experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system. In *Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004)*, Bath, UK, 2004.

[Roberts and Gaizauskas, 2004] Ian Roberts and Robert J. Gaizauskas. Evaluating passage retrieval approaches for question answering. In *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 72–84. Springer, 2004.

[Vicedo *et al.*, 2003] José L. Vicedo, Ruben Izquierdo, Fernando Llopis, and Rafael Muñoz. Question answering in spanish. In *Workshop of the Cross-Lingual Evaluation Forum (CLEF 2003)*, Trondheim, Norway, 2003.

[Voorhees, 1999] Ellen M. Voorhees. The trec-8 question answering track report. In *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 77–82, 1999.

Acknowledgements

We would like to thank ICT EU-India and TEXT-MESS CICYT research projects for partially supporting this work.