JIRS Language-Independent Passage Retrieval System: A Comparative Study

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

Abstract

Passage Retrieval (PR) systems are used as the first step of the actual Question Answering (QA) systems. Usually, PR systems are traditional information retrieval systems which are not oriented to the specific problematic of QA. In fact, these systems only search for the question keywords. We have developed a QA-oriented PR system which searches the question structures in the document collection in order to find the passages with the greatest probability to contain the answer. In this paper, we have carried out a comparative study of our system with other well-known PR models. The experiments show that with our language-independent n-gram model is possible to improve the coverage of the correct answers using natural language questions. The JIRS Distance Density N-gram system has been already adapted to several European languages. At the moment, we have been adapting it also to some of the official Indian languages in order to prove further the independence of the language.

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 problematic 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 including the complete question expression. But 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

[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 this one, to reconstruct its content

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. Whereas Castillo makes 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 a search engine.

In this paper we describe the JAVA Information Retrieval System[3] (JIRS), a QA-oriented Passage Retrieval system. JIRS is able to find the passages using the n-grams of the question and to calculate its similarity with the passage in an effective way. The remainder of this work is structured as it follows. In Sect. 2, we describe the general architecture of the system together with the Distance Density N-gram model of measure to calculate the similarity between passage and question. In Sect. 3 we describe the other PR systems we compared JIRS with. In Sect. 4 we discuss the results of the comparative study. Finally, in Sect. 5 we draw conclusions and future works.

2 Description of the JIRS PR system

JIRS Distance Density N-gram system (Gómez et al., 2005) is a QA-oriented PR system which makes a systematical search of all question n-grams in order to find passages with the greatest probability to contain the correct answer. In order to do it, JIRS uses a traditional PR system as the first step and then searches all possible n-grams of the question in the retrieved passages and rates them due to the number and the weight of n-grams appeared in these passages.

In Figure 1 we can observe the main structure of the system. The *Search Engine* module searches the user question in order to find the passages (i.e., pieces of text) with the question keywords. Every passage returned by the search engine is sorted due to its weight. This passage weight is equal to the sum of all term weights of the question which are found in the passage. The term weight is calculated by:

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

where n_k is the number of passages in which the term t_k occurs and N is the number of system passages.

According to the equation 1, each term has different weight depending on its relevance. For example, stopwords have the least relevance and the terms that appear only once have the most. Therefore, we have that at the top of passage rank returned by the search engine will appear passages with more relevant terms.

With the m most relevant passages[4], the system extracts the 1-grams, 2-grams and so forth to the n-gram (where n is the number of question terms). In parallel, the question n-grams are extracted. Then, the question and passage sets of n-grams are compared using a *Distance Density N-gram* model. This model finds question structures in the passages and gives a higher similarity value to those passages that contain more grouped structures. This similarity value is calculated by:

$$Sim(p,q) = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{\forall x \in P} h(x) \frac{1}{d(x, x_{max})} \quad (2)$$

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:

$$h(x) = \sum_{k=1}^{j} w_k \quad (3)$$

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

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

[4] In previous experiments we have checked that the optimal value of m is between 800 or 1000 for the Spanish CLEF document collection.

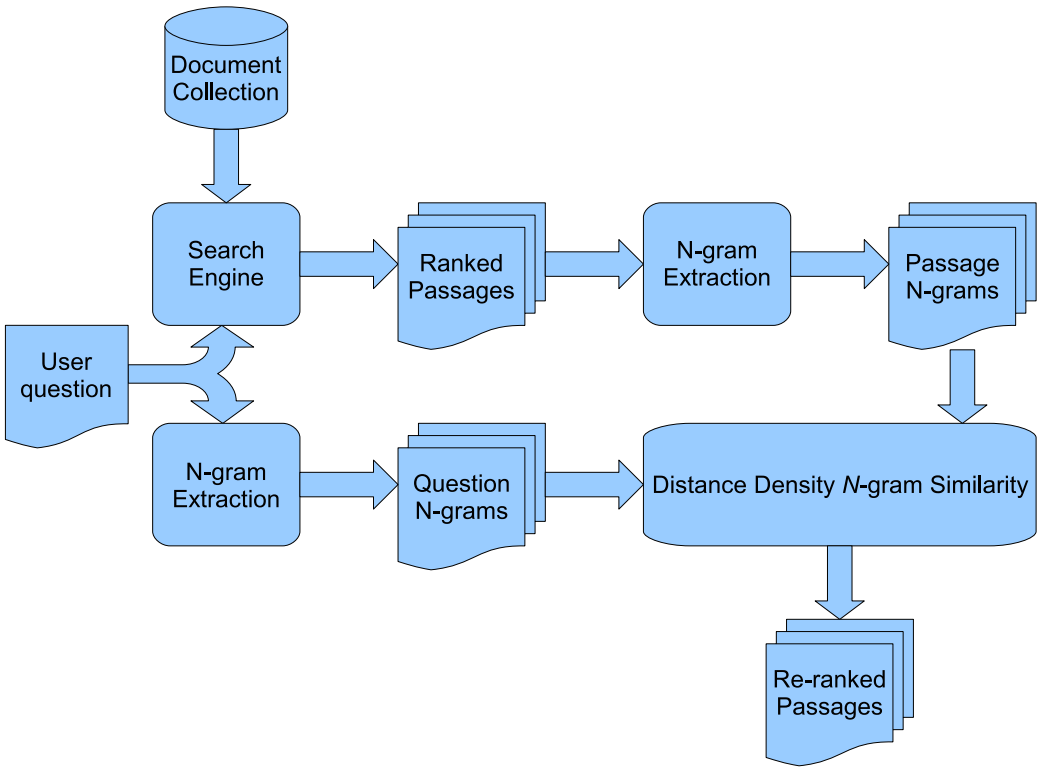


Figure 1: Main structure of JIRS Distance Density N-gram Model

1. 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:

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

Where L is the number of terms between the n-gram x_{max} (x_{max} is the n-gram with the maximum weight calculated in (3)) and the n-gram x of the passage. If there is more than one n-gram x in the passage we choose the closest one. In order to measure the importance degree of the distance factor in the similarity equation, we have introduced the k constant. In previous experiments we have determined that the best value for this is 0.1. The other constants are used to avoid the infinities when L is equal to 0.

$$d(x, x_{max}) = 1 + \ln(1 + L) \qquad (5)$$

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 passage 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 passage has two question n-grams. The greatest n-gram is "*the Croatia*" with a weight of 0.6. The other n-gram question is "*capital of*" with a weight of 0.3. The distance between these n-grams is equal to 7. Therefore, the "*capital*" weight decreases to 0.1 due to the distance factor. If we calculate the similarity for both passages, we obtain the value 0.9 for the first passage and 0.7 for the second one.

In the Distance Density N-gram model, those passages that contain n-grams with more relevant terms have greater weight than others. Therefore, 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 peculiarity of this model is that the similarity value is not affected by the question reformulations. That is, to the n-gram with the expression "*is the capital of Croa-*

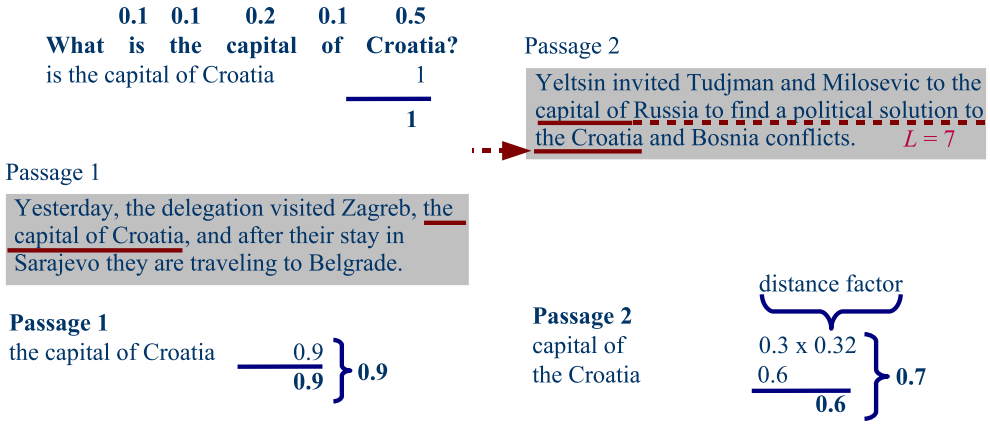


Figure 2: Example of Distance Density N-gram model

tia" will be given the same weight as to "*the capital of Croatia is*" because the distance factor of the n-gram "*is*" to the n-gram "*the capital of Croatia*" is equal to 1. This aspect is very important for languages whose answer expressions are, normally, reformulations of question terms.

The JIRS system has been used in three QA systems that participated in the Cross Language Evaluation Forum 2005 (CLEF)[5] (y Gómez et al., 2006; Gómez et al., 2006). These QA systems have obtained the best results in the Spanish and Italian monolingual tasks and in the English-Spanish and Spanish-English cross-language tasks.

3 Description of Other PR Models

For our study, we implemented 3 different passage retrieval algorithms using as a basis the Lucene[6] search engine. Most of these algorithms are adaptation from the ones selected in the work of Tellex (Tellex et al., 2003), even if the original one were developed for the English language. In our implementations, the first 5000 passages returned by Lucene were re-ranked using the algorithms described below. In some cases, our implementation differs significantly from the original algorithm, due to the fact that some algorithms are not completely independent from the language. The following subsections provide an overview of the used algorithms.

3.1 Improved-MITRE

The original word overlap algorithm presented by Light (Light et al., 2001) simply counts the number of terms a passage has in common with the query. In our variant, the passages are ranked depending also on their length: the weight of each passage (as returned by Lucene) is multiplied by the logarithm of the number of non-whitespace characters in the passage. Although the version described by Light makes use of stemming, we implemented both a stemming and non-stemming version of the algorithm.

3.2 Reduced-MultiText

The original MultiText algorithm (Clarke et al., 2000) is a density-based passage retrieval algorithm that favours short passages containing many terms with high idf values. It makes use of a POS-tagger in order to identify question verbs, which are searched in their stemmed form. Each passage window in the algorithm starts and ends with a query term, and its score is based on the number of query terms in the passage as well as the window size. However, due to the structure of the indices used by Lucene, our implementation returns the whole passage. Our implementation also uses the standard definition of *idf*.

3.3 IR-n-Based

Alicante's passage retrieval algorithm (Llopis-Pascual, 2001) computes the non-length normalised cosine similarity between query terms and the passage. It takes into account the number of appearances of a term in the passage and in the query, along with their idf values.

4 Experiments

The experiments detailed in this paper will be evaluated using a metric known as coverage (for more

[5] http://clef.iei.pi.cnr.it/
[6] http://lucene.apache.org

details see (Roberts and Gaizauskas, 2004)).

Let Q be the question set, P the passage collection, $A_{P,q}$ the subset of P containing correct answers to $q \in Q$, and $R_{P,q,n}$ be the top n ranked documents in P retrieved by the search engine given a question q.

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

$$coverage(Q, P, n) \equiv \frac{|\{q \in Q | R_{P,q,n} \cap A_{P,q} \neq \emptyset\}|}{|Q|} \quad (6)$$

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

Some experiments were carried out on the CLEF Spanish corpus which is composed of documents of the *Agencia EFE (1994/1995)*. The 200 questions which we used are those of the 2005 Spanish QA task. We have used 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 have been used to make two answer collections[7]. 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 the correct answers. The second answer collection is a *lenient* approach and it contains answers with a less strict criterion. For instance, for the question "*What is the FARC?*", a strict criterion would be "*Fuerzas Armadas Revolucionarias de Colombia*" but a lenient criterion would be "*persons in charge of the production of coca and drugs*", "*guerrilla group*" or "*rebellious group*".

In order to evaluate the JIRS Distance Density N-gram model, we must compare our results with those returned by other PR systems. However, the most passage retrieval systems are evaluated with differents parameters and corpora or, simply, they are not evaluated because they are a part of other more complex systems such as QA systems. Therefore, we have decided to implement those passage retrieval algorithms and simulate the same conditions for all PR systems such as the passage size or the number of retrieved passage. In the previous section we have described the different PR algorithms to compare with.

[7]Both sets of answers can be download in http://jirs.dsic.upv.es.

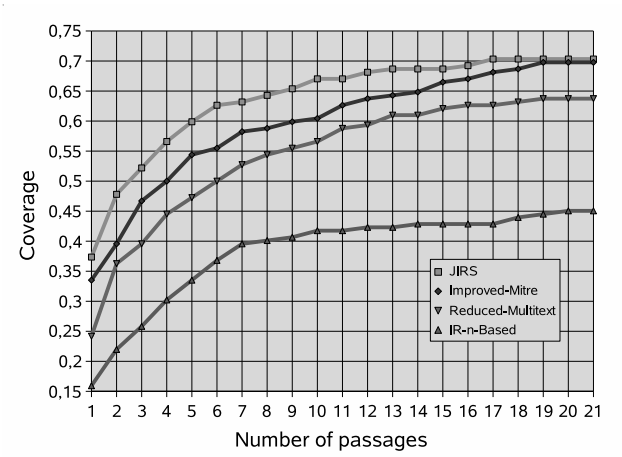
Figure 3 shows that the coverage of different systems (i.e., the number of correct answers for each question) together with the number of retrieved passages using the two differents criterion. Therefore, if we take into account only the first passage returned by these systems in the lenient criterion we have the coverage equal to 0.25, 0.36, 0.49 and 0.55 for the systems IR-n-Based, Reduced-Multitext, Improved-MITRE and JIRS, respectively. As the number of returned passages increases the coverage grows reaching the value of 0.88 for the Improved-MITRE and 0.89 for the JIRS systems.

In Figure 3 we can observe how the Distance Density N-gram model improves the coverage with regard to other implemented models. The IR-n-Based is the system which obtains the lowest performance. Reduced-Multitext model has better results than the IR-n-Based model but it is about 10 points below the Improved-Mitre one. The Improved-Mitre model results are the most similar to JIRS: even if it obtains enough coverage at the 20 passages, this model is worse than JIRS in the first retrieved passages.

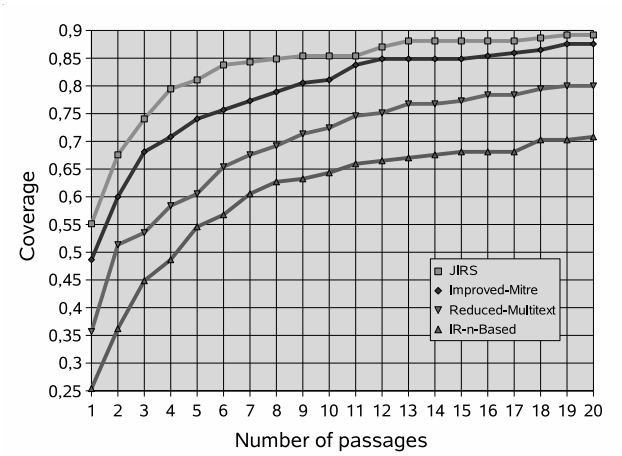
A similar behaviour is observed in the strict criterion figure where the results of the strict task are represented.

When we consider the strict criterion, the coverage obtained by all systems is lower, because of the smaller number of possible answers available for each question. In this task the difference between the IR-n-Based and the other models has increased. The rest of systems have less differences. Therefore, the coverage of IR-n-Based and Reduced-Multitext at the 20th passage is equal to 0.45 and 0.64 respectively, and it is equal to 0.7 for the Improved-MITRE and the JIRS systems, even if JIRS continue having the highest coverage in the first passages.

In both figures it can be observed that JIRS obtained better results than the ones obtained by the other systems; notably, it can achieve a good level of coverage just with the first passages. Besides, the similar behaviour of the lenient and strict tasks indicates that our system is independent of the evaluation criterion.



(a) Strict evaluation



(b) Lenient evaluation

Figure 3: Comparison of the different PR models by means of the lenient criterion

5 Conclusions and Further Work

Our QA-oriented PR system makes a better use of the redundancy bearing in mind all the possible reformulations of the question efficiently running the search engine with just one question.

According to the results represented in this paper, we can observe that the JIRS model notably improves the coverage, mainly in the first passages. The matter is that IR-n-Based and Improved-MITRE use only keywords as the main information. Whereas the Reduced-Multitext model uses, also, lexical categories, they can contain mistakes due to POS ambiguity errors. JIRS not only searches for keywords, it also uses morphologic information to increase the probability of occurrence of the correct answer in the retrieved passages. Moreover, the distance-based model of JIRS evaluates the density of query n-grams occurred in the passages. Therefore, the passages that contain n-grams composed by relevant keywords, and the n-grams themselves are distributed narrowly in the passage, have a greater weight than other ones. It is worth noting that the JIRS Distance Density N-gram model success in returning the correct answer in the first passages. For instance, in the case of 4 retrieved passages the coverage exceeds the results of the other systems by more than 9%.

Although the differences with respect to the Improved-MITRE are not very significative at 20th passage, we believe that the use of the JIRS Distance Density N-gram system is better for two main reasons: it is language-independent and it has a high coverage in the first passages. the main drawback of our system is that its results depend on the size of the corpus of documents. If the corpus is too small, then it is not possible to obtain enough redundancy for the JIRS Distance Density N-Gram model. This problem can be addressed by means of query reformulations or using bigger

text corpora such as the Internet.

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. In any language with few differences between the question and the answer sentences, our system should work very well. For this reason, we are investigating this hypothesis by adapting our PR retrieval system to many of the official European and also Indian languages.

References

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

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

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

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

Charles L. A. Clarke, Gordon V. Cormack, D. I. E. Kisman, and Thomas R. Lynam. 2000. Question answering by passage selection (multitext experiments for trec-9). In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*.

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

Alejandro Del-Castillo-Escobedo, Manuel Montes-Gómez, and Luis Villaseñor-Pineda. 2004. 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.

José Manuel Gómez, Empar Bisbal-Asensi, Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. 2005. Monolingual and cross-language qa using a qa-oriented passage retrieval system. In *Working Notes for the CLEF 2005 Workshop; to Be Published in Springer Verlag Special Issue on "Cross Language Evalutaion Forum 2005" as "QUASAR: The Question Answering System of the Universidad Politecnica de Valencia", in Press*, Vienna, Austria.

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

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

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

Marc Light, Gideon S. Mann, Ellen Riloff, and Eric Breck. 2001. Analyses for elucidating current question answering technology. In *Journal of Natural Language Engineering*.

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

Fernando Llopis-Pascual. 2001. *IR-N: Un Sistema de Recuperación de Información Basado en Pasajes*. Phd. Thesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, Spain. .

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

Günter Neumann and Bogdan Sacaleanu. 2004. 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.

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

Stefanie Tellex, Boris Katz, Jimmy J. Lin, Aaron Fernandes, and Gregory Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR*, pages 41–47.

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

Manuel Montes y Gómez, Luis Villase nor Pineda, Manuel Pérez-Couti no, José Manuel Gómez-Soriano, Emilio Sanchis-Arnal, and Paolo Rosso, 2006. *A Full Data-Driven System for Multiple Language Question Answering*, volume 4022 of *Lecture Notes in Computer Science*, pages 439–448. Springer-Verlag GmbH, Vienna, Austria.

Acknowledgements

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