

# Two Approaches for Multilingual Question Answering: Merging Passages vs. Merging Answers

Rita M. Aceves-Pérez<sup>†</sup>, Manuel Montes-y-Gómez<sup>†</sup>,  
Luis Villaseñor-Pineda<sup>†</sup>, and L. Alfonso Ureña-López<sup>‡</sup>

## Abstract

One major problem in multilingual Question Answering (QA) is the integration of information obtained from different languages into one single ranked list. This paper proposes two different architectures to overcome this problem. The first one performs the information merging at passage level, whereas the second does it at answer level. In both cases, we applied a set of traditional merging strategies from cross-lingual information retrieval. Experimental results evidence the appropriateness of these merging strategies for the task of multilingual QA, as well as the advantages of multilingual QA over the traditional monolingual approach.

**Keywords:** multilingual question answering, cross-lingual information retrieval, information merging.

## 1 Introduction

Question Answering (QA) has become a promising research field whose aim is to provide more natural access to textual information than traditional document retrieval techniques [Laurent et al. 2006]. In essence, a QA system is a kind of search engine that responds to natural language questions with concise and precise answers. For instance, given the question “Where is the

---

<sup>†</sup> Laboratory of Language Technologies, National Institute of Astrophysics, Optics and Electronics (INAOE). Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, Mexico.

Email: {rmaceves, mmontesg, villasen}@ccc.inaoep.mx

<sup>‡</sup> Department of Computer Science, University of Jaén. Campus Las Lagunillas s/n, Edif D3, Jaén, Spain.

Email: laurena@ujaen.es

Popocatepetl Volcano located?”, a QA system has to respond “Mexico”, instead of returning a list of related documents to the volcano.

At present, due to the internet explosion and the existence of several multicultural communities, one of the major challenges that faces this kind of systems is the *multilinguality*. In a multilingual scenario, it is expected that QA systems will be able to: (i) answer questions formulated in several languages, and (ii) look for answers in a number of collections in different languages.

There are two recognizable kinds of QA systems that allow to manage information in various languages: cross-lingual QA systems and, the strictly speaking, *multilingual QA systems*. The formers address the situation where questions are formulated in a different language from that of the (single) document collection. The others, in contrast, perform the search over two or more document collections in different languages.

It is important to mention that both kinds of systems have some advantages over standard monolingual QA. They mainly allow users to access more information in an easier and faster way than monolingual systems. However, they also introduce additional issues due to the language barrier.

Generally speaking, a multilingual QA system can be described as an *ensemble* of several monolingual systems, where each one works on a different –monolingual– document collection. Under this schema, two additional tasks are required: first, the translation of incoming questions to all target languages, and second, the combination of relevant information extracted from different languages.

The first problem, namely, the translation of questions from one language to another, has been widely studied in the context of cross-language QA [Aceves-Pérez et al. 2007; Neumann et al. 2005; Rosso et al. 2007; Sutcliffe et al. 2005]. In contrast, the second task, the merging of information obtained from different languages, has not been specifically addressed in QA.

Nevertheless, it is important to mention that there is significant work on combining capacities from several monolingual QA systems [Chu-Carroll et al. 2003; Jijkoun et al. 2004; Sangoi-Pizzato et al. 2005], as well as on merging multilingual lists of documents for cross-lingual information retrieval applications [Lin et al. 2002; Savoy et al. 2004].

In line with these previous works, in this paper we propose *two different architectures for multilingual question answering*. These architectures differ from each other by the way they handle the combination of multilingual information. Mainly, they take advantage of the pipeline architecture of monolingual QA systems (which includes three main modules, one for question classification, one for passage retrieval, and one for answer extraction) to achieve this combination at two different stages: after the passage retrieval module by mixing together the sets of recovered passages, or after the answer extraction module by directly combining all extracted answers. In other words, our first architecture performs the combination at *passage level*, whereas the second approach does it at *answer level*. In both cases, we applied a set of well-known strategies for information merging from cross-lingual information retrieval, specifically, Round Robin, Raw Score Value (RSV), CombSUM, and CombMNZ [Lee et al. 1997; Lin et al. 2002; Savoy et al. 2004].

The contributions of this paper are two-fold. On the one hand, it represents –in our knowledge– the first attempt for doing “multilingual” QA. In particular, it proposes and compares two initial solutions to the problem of multilingual information merging in QA. On the other hand, this paper also provides some insights on the use of traditional ranking strategies from cross-language information retrieval into the context of multilingual QA.

The rest of the paper is organized as follows. Section 2 describes some previous works on information merging. Section 3 presents the proposed architectures for multilingual QA. Section 4 describes the procedures for passage and answer merging. Section 5 shows some experimental results. Finally, section 6 presents our conclusions and outlines future work.

## 2 Related Work

As we previously mentioned, a multilingual QA system has to consider, besides the traditional modules for monolingual QA, stages for question translation and information merging.

The problem of question translation has already been widely studied. Most current approaches rest on the idea of combining capacities of several translation machines. They mainly consider the selection of the best instance from a given set of translations [Aceves-Pérez et al. 2007; Rosso et al. 2007] as well as the construction of a new question reformulation by gathering terms from all of them [Neumann et al. 2005; Sutcliffe et al. 2005; Aceves-Pérez et al. 2007].

On the other hand, the problem of information merging in multilingual QA has not been already addressed. However, there is some relevant related work on constructing ensembles of monolingual QA systems. For instance, [Jijkoun et al. 2004] proposes a method that performs a number of sequential searches over different document collections. At each iteration, this method filters out or confirms the answers found in the previous step. [Chu-Carroll et al. 2003] describes a method that applies a general ranking over the five-top answers obtained from different collections. They use a ranking function that is inspired in the well-known RSV technique from cross-language information retrieval. Finally, [Sangoi-Pizzato et al. 2005] uses various search engines in order to extract from the Web a set of candidate answers for a given question. It also applies a general ranking over the extracted answers; nevertheless, in this case the ranking function is based on the confidence of search engines instead that on the redundancy of individual answers.

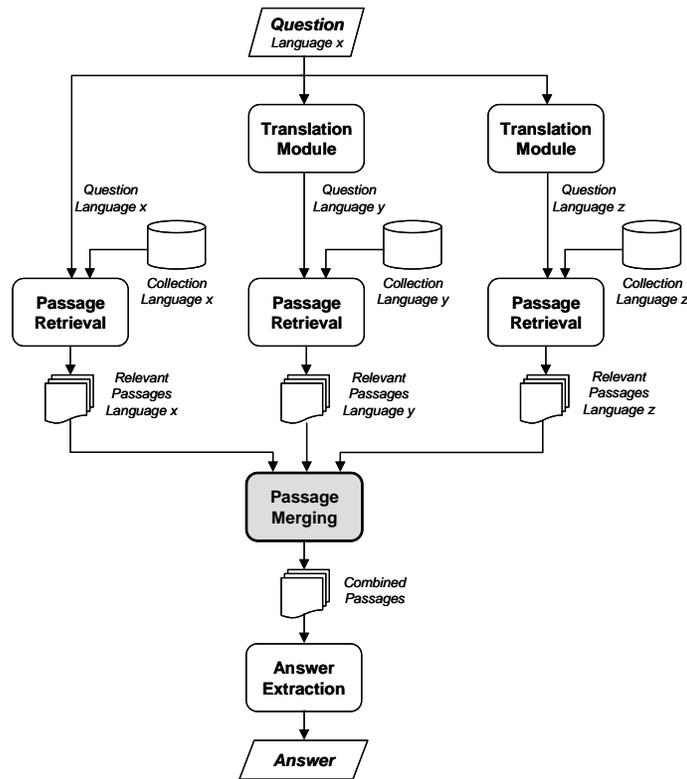
Our proposal mainly differs from previous methods in that it not only considers the integration of answers but also takes into account the combination of passages. That is, it also proposes a method that carries out the information merging at an internal stage of the QA process. The proposed merging approach is similar in spirit to [Chu-Carroll et al. 2003; Sangoi-

Pizzato et al. 2005] in that it also applies a general ranking over the information extracted from different languages. Like [Chu-Carroll et al. 2003], it uses the RSV ranking function, although it also applies other traditional ranking strategies such as Round Robin, CombSUM and CombMNZ.

### **3 Two Architectures for Multilingual QA**

The traditional architecture of a monolingual QA system considers three basic modules: (i) question classification, where the type of expected answer is determined, (ii) passage retrieval, where the passages with the greatest probability to contain the answer are obtained from the target document collection, and (iii) answer extraction, where candidate answers are ranked and the final answer recommendation of the system is produced. In addition, a multilingual QA system must include two other modules, one for question translation and another for information merging. The purpose of the first module is to translate the input question to all target languages, whereas the second module is intended to integrate the information extracted from these languages into one single ranked list.

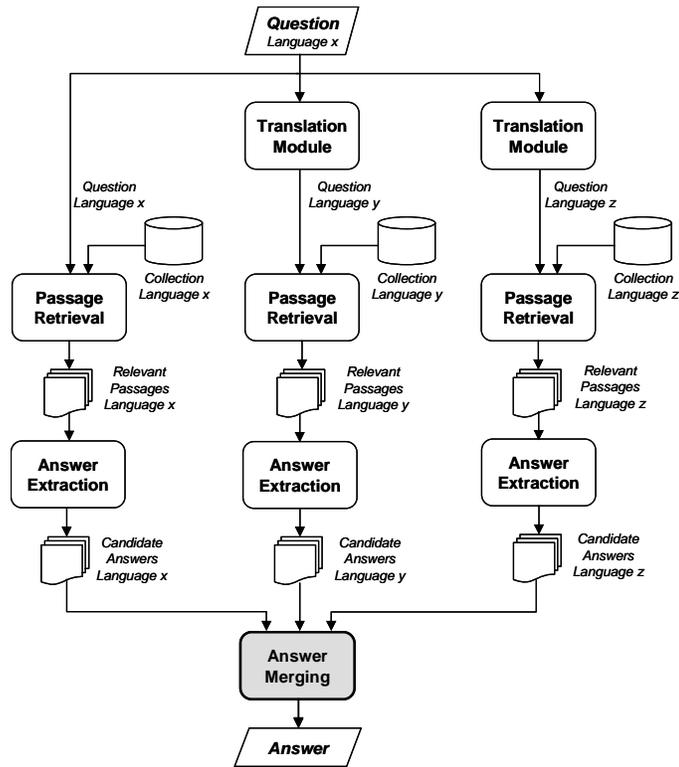
Figures 1 and 2 show two different architectures for multilingual QA. For the sake of simplicity, in both cases we do not consider the module for question classification. On one hand, Figure 1 shows a multilingual QA architecture that does the information merging at *passage level*. The idea of this approach is to perform in parallel the recovery of relevant passages from all collections (i.e., from all different languages), then integrate these passages into one single ranked list, and finally, extract the answer from the combined set of passages. On the contrary, Figure 2 illustrates an architecture that achieves the information merging at *answer level*. In this case, the idea is to perform the complete QA process independently in all languages, and after that, integrate the sets of answers into one single ranked list.



*Figure 1. Multilingual QA based on passage merging*

It is important to mention that merging processes normally rely on the translation of information to a common language. This translation is required for some merging strategies in order to be able to compare and rank the passages and answers extracted from different languages.

The two proposed architectures have different advantages and disadvantages. For instance, doing the information merging at passage level commonly allows obtaining better translations for named entities (possible answers) since they are immerse in an extended context. On the other hand, doing the merging at answer level has the advantage of a clear (unambiguous) comparison of the multilingual information. In other words, comparing two answers (named entities) is a straightforward step, whereas comparing two passages requires the definition of a similarity measure and the determination of a criterion about how similar two different passages



*Figure 2. Multilingual QA based on answer merging*

should be in order to be considered as equal. This previous problem is not present in monolingual QA ensembles, since in that case all individual QA systems search on the same document collection.

The following section introduces some of the most popular information merging strategies used in the task of cross-lingual information retrieval. It also describes the way these strategies are used within the proposed architectures for integrating passages and answers.

## 4 Merging Passages and Answers

### 4.1 Merging Strategies

Integrating information retrieved from different document collections or by different search engines is a longstanding problem in information retrieval. Researchers in this field have

proposed several strategies for information merging; traditional ones are: Round Robin, RSV (Raw Score Value), CombSUM, and CombMNZ [Lee et al. 1997; Lin et al. 2002]. However, recently more sophisticated strategies have been proposed, such as the 2-step RSV [Martínez-Santiago et al. 2006], and the Z-score value [Savoy et al. 2004].

In this work we mainly study the application of traditional merging strategies in the context of multilingual QA. The following paragraphs give a brief description of these strategies.

**Round Robin.** The retrieved information (passages or answers in our case) from different languages is interleaved according to its original monolingual rank. In other words, this strategy takes one result in turn from each individual list and alternates them in order to construct the final merged output. The hypothesis underlying this strategy is the homogeneous distribution of relevant information across all languages. In our particular case, as described in Table 1, this restriction was fulfilled for almost 60% of test questions.

**Raw Score Value (RSV).** This strategy sorts all results (passages or answers) by their original score computed independently from each monolingual collection. Different to Round Robin, this approach is based on the assumption that scores across different collections are comparable. Therefore, this method tends to work well when different collections are searched by the same or very similar methods. In our experiments (refer to section 5), this condition was fully satisfied since it was applied the same QA system for all languages.

**CombSUM.** In this strategy, the result scores from each language are initially (min-max) normalized. Afterward, the scores of duplicated results occurring in multiple collections are summed. In particular, we considered the implementation proposed by [Lee et al. 1997]: we assigned a score of  $21-i$  to the  $i$ -th ranked result from the top 20 of each language, this way, the top passage or answer was scored 20, the second one was scored 19, and so on. Any result not ranked in the top 20 was scored as 0. Finally, we added scores of duplicated results for all different monolingual runs and ranked these results in accordance to their new joint score. For

instance, if an answer is ranked 3<sup>rd</sup> for one language, 10<sup>th</sup> for other one, and does not exist in a third language, then its score is  $(21-3) + (21-10) + 0 = 29$ .

**CombMNZ.** It is based on the same normalization that CombSUM, but also attempts to account for the value of multiple evidence by multiplying the sum of the scores (CombSUM-value) of a result by the number of monolingual collections in which it occurs. Therefore, it can be said that CombSUM is equivalent to an averaging, whereas CombMNZ is equivalent to a weighted averaging. Using the same example as for the CombSUM strategy, the answer's score is in this case  $2 \times ((21-3) + (21-10) + 0) = 58$ .

It is important to point out that Round Robin and RSV strategies take advantage of the complementarity among collections (when answers are extracted from only one language), whereas ComSUM and CombMNZ also take into account the redundancies of answers (the repeated occurrence of an answer in several languages).

## 4.2 Merging Procedures

Given several sets of relevant passages obtained from different languages, the procedure for passage merging considers the following two basic steps:

1. Translate all passages to one common language. This translation can be done by means of any translation method or online translation machine. However, we suggest translating all passages into the original question's language in order to avoid translation errors in at least one passage set.

It is important to clarify that translation is only required by the CombSUM and CombMNZ strategies. Nevertheless, all passages should be translated to one common language before entering the answer extraction module.

2. Combine the sets of passages according to a selected merging strategy. In the case of using the Round Robin or RSV approaches, the combination of passages is straightforward. In contrast, when applying CombSUM or CombMNZ, it is necessary to determine the

occurrence of a given passage in two or more collections. Given that it is practically impossible to obtain exactly the same passage from two different collections, it is necessary to define a criterion about how similar two different passages should be in order to be considered as equal. In particular, we measure the similarity of two passages by the Jaccard function (calculated as the cardinality of their vocabulary intersection divided by the cardinality of their vocabulary union) and consider them as equal only if their similarity is greater than a given specified threshold (empirically we set the threshold value to 0.5).

The procedure for answer merging is practically the same as that for passage merging. It also includes one step for answer translation and another step for answer combination. However, the combination of answers is much simpler than the combination of passages, since they are directly comparables. In this case, the application of all merging strategies is straightforward.

## 5 Experimental Evaluation

### 5.1 Experimental Setup

The following paragraphs describe the data and tools used in the experiments.

**Languages.** We considered three different languages: Spanish, Italian, and French.

**Search Collections.** We used the document sets from the QA@CLEF evaluation forum. In particular, the Spanish collection consists of 454,045 news documents, the Italian set has 157,558, and the French one contains 129,806.

**Test questions.** We selected a subset of 170 factoid questions from the MultiEight corpus of CLEF. From all these questions at least one monolingual QA system could extract the correct answer. Table 1 shows answer's distributions across all languages.

It is important to mention that this set of questions covers all types of currently-evaluated factoid questions, and that therefore it makes possible to formulate some accurate conclusions about the appropriateness of the proposed architectures.

*Table 1. Distribution of questions by source language*

	<i>Answers in:</i>						
	<b>SP</b>	<b>FR</b>	<b>IT</b>	<b>SP, FR</b>	<b>SP, IT</b>	<b>FR, IT</b>	<b>SP, FR, IT</b>
<i>Questions</i>	37	21	15	20	25	23	29
<i>Percentage</i>	21%	12%	9%	12%	15%	14%	17%

**Monolingual QA System.** We used the passage retrieval and answer extraction components of the TOVA question answering system [Montes-y-Gómez et al. 2005]. Its selection was mainly supported on its competence to deal with all considered languages. Indeed, it obtained the best precision rate for Italian and the second best for both Spanish and French in the CLEF-2005 evaluation exercise.

**Translation Machine.** The translation of passages and answers was done using the Systran online translation machine ([www.systranbox.com](http://www.systranbox.com)). On the other hand, questions were manually translated in order to avoid mistakes at early stages and therefore focus the evaluation on the merging phase.

**Merging strategies.** As we mentioned in the previous section, we applied four traditional merging strategies, namely, Round Robin, RSV, CombSUM and CombMNZ.

**Evaluation Measure.** In all experiments we used the precision as evaluation measure. It indicates the general proportion of correctly answered questions. In order to enhance the analysis of results we show the precision at one, three, and five positions.

**Baseline.** We decided to use as baseline the results from the best monolingual system (the Spanish system in this case). In this way, it is possible to reach conclusions about the advantages of multilingual QA over the standard monolingual approach.

## 5.2 Experimental Results

The objectives of the experiments were two fold: first, to compare the performance of both architectures; and second, to study the applicability and usefulness of traditional merging strategies in the problem of multilingual QA. Additionally, these experiments allowed to analyze the advantages of multilingual QA over the traditional monolingual approach.

The first experiment considered the information merging at passage level. In this case, the passages obtained from different languages were combined, and the 20 top-ranked were delivered to the answer extraction module. Table 2 shows the precision results obtained using all merging strategies as well as the precision rates of the best monolingual run.

*Table 2. Precision results of the passage merging approach*

<b>Merging Strategy</b>	<b>Precision at:</b>		
	<b>1<sup>st</sup></b>	<b>3<sup>rd</sup></b>	<b>5<sup>th</sup></b>
Round Robin	0.41	0.57	0.65
RSV	<b>0.45</b>	<b>0.65</b>	<b>0.66</b>
CombSUM	0.40	0.54	0.64
CombMNZ	0.40	0.54	0.63
Best Monolingual	0.45	0.57	0.64

From Table 2, it is clear that merging strategies relying on the complementarity of information (such as Round Robin and RSV) obtain better results than those also considering its redundancy (e.g. CombSUM and CombMNZ). We hypothesize that this behavior was mainly produced by three different factors: (i) the impact of translation errors on the CombSUM and CombMNZ strategies<sup>1</sup>; (ii) the complexity of assessing the redundancy of passages, i.e., the complexity of correctly deciding whether two different passages should be considered as equal, and (iii) the large number of questions (42%) that have answer in just one language.

---

<sup>1</sup> We do not have an exact estimation of the translation errors for this task, but we suppose they are very abundant. This supposition is based on current reports from cross-lingual QA (Vallin et al. 2005) which indicate severe reductions –as high as 60%– in precision results as consequence of unsatisfactory question translations.

The second experiment achieved information merging at answer level. In this experiment, we considered the 10 top-ranked answers from each monolingual QA system. Table 3 shows the results obtained using all different merging strategies.

*Table 3. Precision results of the answer merging approach*

<b>Merging Strategy</b>	<i>Precision at:</i>		
	<b>1<sup>st</sup></b>	<b>3<sup>rd</sup></b>	<b>5<sup>th</sup></b>
Round Robin	<b>0.45</b>	<b>0.68</b>	0.74
RSV	0.44	0.61	0.69
CombSUM	0.42	0.66	<b>0.75</b>
CombMNZ	0.42	0.62	0.70
Best Monolingual	0.45	0.57	0.64

The results of Table 3 are encouraging. They show that all merging strategies achieved high performance levels, improving baseline results at the third and fifth positions by more than 7% and 8% respectively. Once again these results indicate that simple strategies outperformed complex ones. However, they do not necessarily mean that Round Robin and RSV are better than CombSum and CombMNZ, instead they only express that the formers are less sensible to translation errors.

Comparing the results of both architectures, it is easy to observe that merging answers allowed obtaining better precision rates than merging passages. It seems that this situation is because the combination of answers is easier than the combination of passages, and therefore, the first one allows to better taking advantage of both the complementarity as well as the redundancy of information. This phenomenon is more evident in the performance of CombSUM and CombMNZ; in the case of passage merging, their results were always below the baseline, and were –in average– 6% below the best precision rate, whereas in answer merging they were only 3% below the best result.

In addition, the fact that the RSV was the best strategy for passage merging and Round Robin for answer merging evidence, on one hand, the pertinence of the passage scores against the low confidence of the answer scores, and on the other hand, the homogeneous distribution of the answers in all languages (from Table 1: 65% of the questions has an answer –at the first 20 positions– in Spanish, 55% in French and 55% in Italian).

## 6 Conclusions

The problem of cross-lingual QA has been widely studied; nevertheless –in our knowledge– there are not specific solutions to the related problem of multilingual QA. Particularly, this paper focused on this new direction. It proposed *two different architectures for multilingual QA*. One of them performs information merging at passage level, whereas the other does it at answer level.

A secondary contribution of our work, but not necessarily less important, is the study of the *usefulness of traditional ranking strategies* from cross-language information retrieval into the context of multilingual QA.

The presented experimental results allowed to reach the following conclusions:

1. *A multilingual QA system may help respond a large number of questions than a traditional monolingual QA system.* Considering that practical QA systems supply lists of candidate answers instead of isolated responses, our results demonstrated that, using a simple multilingual QA approach, it was possible to answer up to 10% more questions than using a traditional monolingual system.
2. *Merging answers seems to be more convenient than merging passages.* This assertion is mainly supported on the fact that it is more difficult to observe and compute the information redundancy at passage level than at answer level. In addition, the results of passage merging will be inevitably affected by the (quality of the) answer extraction module, whereas the results of answer merging are the actual output.

3. Translation errors directly affect the performance of some merging strategies. It seems that merging strategies such as CombSUM and CombMNZ are more relevant than the rest (simple ones such as Round Robin and RSV). However, our results demonstrate that they are more sensible to translation mistakes.

Finally, in order to improve the results of multilingual QA we plan to investigate the following issues:

- Different criteria to evaluate the similarity between passages. In particular, we consider that this action can have an important influence on the performance of strategies based on the information redundancy, such as CombSUM and CombMNZ.
- Ensemble methods for improving the translation of passages and answers. We plan to work with methods that combine the capacities from several translation machines, by selecting the best instance from a given set of translations, or by constructing a new translation reformulation by gathering terms from all of them.
- New merging strategies. In particular we are considering applying graph and probabilistic based ranking techniques. We believe these kinds of techniques will help developing more robust multilingual merging strategies.

**Acknowledgements.** This work was done under partial support of CONACYT (Project Grant 43990), SNI-Mexico, and the Human Language Technologies Laboratory at INAOE. We also want to thanks to the CLEF organization as well as the EFE agency for the resources provided.

## References

Aceves-Pérez R., Montes-y-Gómez M., Villaseñor-Pineda L., “Enhancing Cross-Language Question Answering by Combining Multiple Question Translations”, *CICLing-2007, Lecture Notes in Artificial Intelligence 4394*, Springer 2007.

- Chu-Carroll J., Czuba K., Prager A. J., Ittycheriah A., “In Question Answering, Two Heads are Better than One”, *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 1, Canada, 2003.
- Jijkoun V., Mishne G., Rijke M., Schlobach S., Ahn D., Muller K., “The University of Amsterdam at QA@CLEF 2004”, *Working Notes of CLEF 2004*, Bath, UK, 2004.
- Laurent D., Séguéla P., Nègre S., “QA better than IR?”, *Workshop on Multilingual Question Answering MLQA-2006*, Trento, Italy, 2006.
- Lee J., “Analysis of Multiple Evidence Combination”, *20<sup>th</sup> annual ACM SIGIR Conference*, 1997.
- Lin W. C., Chen H. H., “Merging Mechanisms in Multilingual Information Retrieval”, *Working Notes CLEF 2002*, Rome, Italy, 2002.
- Martínez-Santiago F., Ureña-López L. A., Martín-Valdivia M., “A merging strategy proposal: The 2-step retrieval status value method”, *Information Retrieval*, vol. 9, no. 1, 2006.
- Montes-y-Gómez M., Villaseñor-Pineda L., Pérez-Coutiño M., Gómez-Soriano J. M., Sanchis-Arnal E., Rosso P., “INAOE-UPV Joint Participation in CLEF 2005: Experiments in Monolingual Question Answering”, *Working Notes CLEF 2005*. Vienna, Austria, 2005.
- Neumann G., Sacaleanu B., “DFKI’s LT-lab at the CLEF 2005 Multiple Language Question Answering Track”, *Working Notes CLEF 2005*, Vienna, Austria, 2005.
- Rosso P., Buscaldi D., Iskra M., “Web-based Selection of Optimal Translations of Short Queries”, *Procesamiento de Lenguaje Natural*, Num. 38, 2007.
- Sangoi-Pizzato L. A., Molla-Aliod D., “Extracting Exact Answers using a Meta Question Answering System”, *Australasian Language Technology Workshop*, Australia, 2005.
- Savoy J., Berger P. Y., “Selection and Merging Strategies for Multilingual Information Retrieval”, *Working Notes of CLEF 2004*, Bath, UK, 2004.
- Sutcliffe R., Mulcahy M., Gabbay I., O’Gorman A., White K., Slatter D., “Cross-Language French-English Question Answering using the DLT System at CLEF 2005”, *Working Notes CLEF 2005*, Vienna, Austria, 2005.
- Vallin A., Giampiccolo D., Aunimo L., Ayache C., Osenova P., Peñas A., de Rijke M., Sacaleanu B., Santos D., Sutcliffe R. Overview of the CLEF 2005 Multilingual Question Answering Track, *Working Notes CLEF 2005*, Vienna, Austria, 2005.