Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 27-40

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

Rita M. Aceves-Pérez*, Manuel Montes-y-Gómez*,

Luis Villaseñor-Pineda*, and L. Alfonso Ureña-López+

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

* Laboratory of Language Technologies, National Institute of Astrophysics, Optics and Electronics (INAOE). Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, Mexico.
Tel.: +52-222-2663100 ext: 8218 Fax: +52-222-2663152.
The author for correspondence is Manuel Montes-y-Gómez.
Email: mmontesg@inaoep.mx

+ Department of Computer Science, University of Jaén. Campus Las Lagunillas s/n, Edif D3, Jaén, Spain

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

At present, due to the internet explosion and the existence of several multicultural communities, one of the major challenges to face this kind of system is *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 management of information in various languages: cross-lingual QA systems and, strictly speaking, *multilingual QA systems*. The former addresses a situation where questions are formulated in a different language from that of the (single) document collection. The other, in contrast, performs 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 into 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; Ahn *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 – to 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. In addition, 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, in addition to 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 yet been addressed. However, there is some relevant related work on constructing ensembles of monolingual QA systems. For instance, [Ahn *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

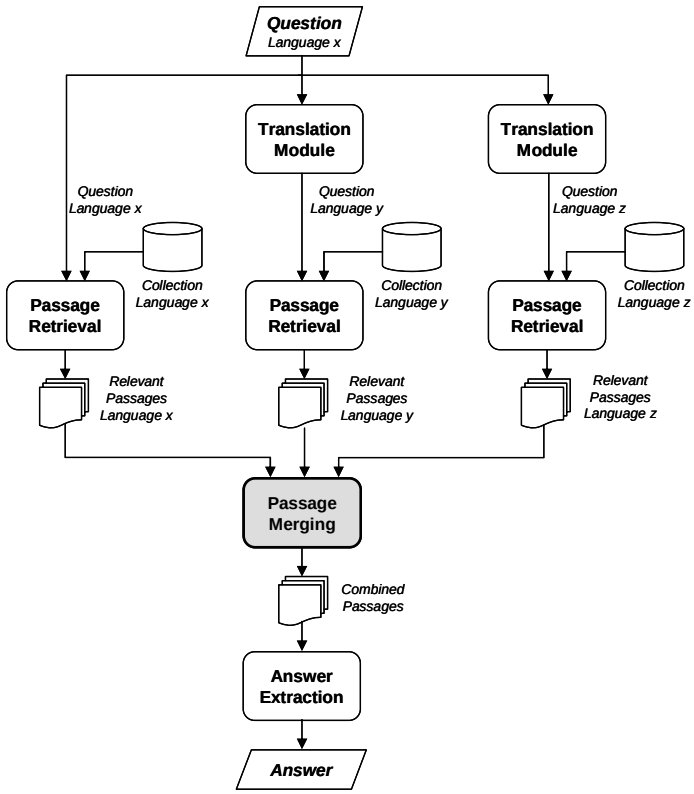

Figure 1. Multilingual QA based on passage merging

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] and 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

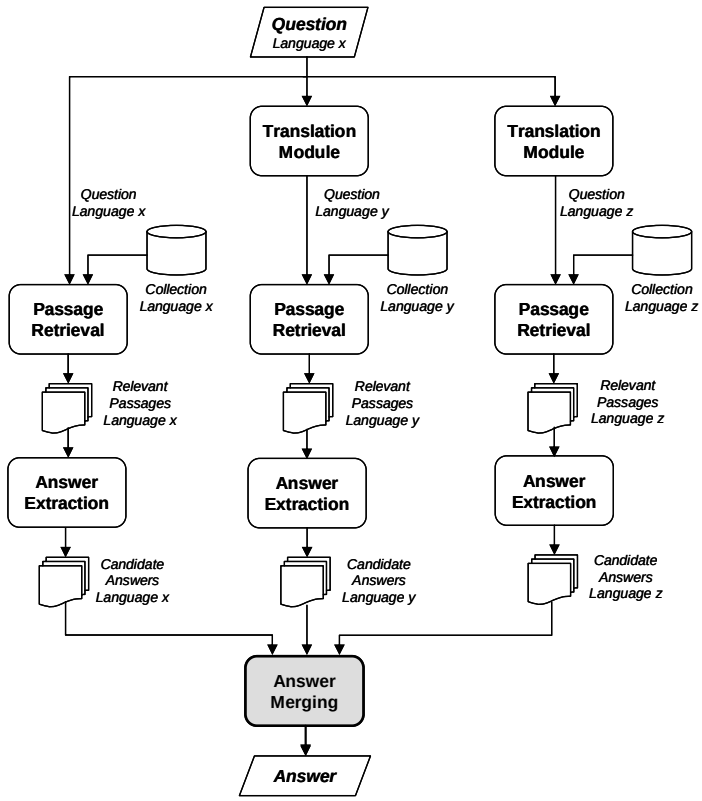

Figure 2. Multilingual QA based on answer merging

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 the 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 then 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.

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 immersed 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 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, more sophisticated strategies have been proposed recently, 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 (in this case, passages or answers) 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. Differing from 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 3rd for one language, 10th 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 as 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 averaging, whereas CombMNZ is equivalent to weighted averaging. Using the same example as for the CombSUM strategy, the answer's score is in this case 2 × ((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 into 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 comparable. In this case, the application of all merging strategies is straightforward.

5. 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.

Table 1. Distribution of questions by source language

	Answers in:						
	SP	**FR**	**IT**	**SP, FR**	**SP, IT**	**FR, IT**	**SP, FR, IT**
Questions	37	21	15	20	25	23	29
Percentage	21%	12%	9%	12%	15%	14%	17%

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

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 by its competence in dealing with all the 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). 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 the 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 the results from the best monolingual system (the Spanish system in this case) as a baseline. 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 twofold: 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 us to analyze the advantages of multilingual QA over the traditional monolingual approach.

The first experiment considered 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.

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[1]; (*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 an answer in just one language.

[1] 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 a consequence of unsatisfactory question translations.

Table 2. Precision results of the passage merging approach

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

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

Merging Strategy	*Precision at*:		
	1st	**3rd**	**5th**
Round Robin	**0.45**	**0.68**	0.74
RSV	0.44	0.61	0.69
CombSUM	0.42	0.66	**0.75**
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 former methods are less sensitive to translation errors.

Comparing the results of both architectures, it is easy to observe that merging answers obtained better precision rates than merging passages. It seems that this situation is because the combination of answers is easier than the combination of passages; 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 – on average – 6% below the best precision rate, whereas, in answer merging, they were only 3% below the best result.

In addition, the fact that RSV was the best strategy for passage merging and Round Robin for answer merging shows, on the 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 – to our knowledge – there are no specific solutions to the related problem of multilingual QA. 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 us to reach the following conclusions:

A multilingual QA system may help respond to a larger 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.

Merging answers seems to be more convenient than merging passages. This assertion is mainly supported by 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 inevitably be affected by the (quality of the) answer extraction module, whereas the results of answer merging are the actual output.

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 sensitive to translation mistakes.

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

1. Using 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.
2. Using ensemble methods for improving the translation of passages and answers. We plan to work with methods that combine the capacities of 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.

3. Using new merging strategies. In particular, we are considering applying graph and probabilistic based ranking techniques. We believe these kinds of techniques will help develop 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., M. Montes-y-Gómez, and L. Villaseñor-Pineda, "Enhancing Cross-Language Question Answering by Combining Multiple Question Translations," In *Proceedings of the 8th Internactional Conference in Computational Linguistics and Intelligent Text Processing CICLing-2007*, 2007, Mexico City, Mexico, pp. 485-493.

Ahn, D., V. Jijkoun, K. Müller, M. de Rijke, S. Schlobach, and G. Mishne, "Making Stone Soup: Evaluating a Recall-Oriented Multi-stream Question Answering System for Dutch," In *Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum CLEF 2004*, 2004, Bath, UK, pp. 423-434.

Chu-Carroll, J., K. Czuba, A. J. Prager, and A. Ittycheriah, "In Question Answering, Two Heads are Better than One," In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology HLT-NAACL 2003*, 2003, Edmonton, Canada, pp. 24-31.

Laurent, D., P. Séguéla, and S. Nègre, "QA better than IR?," In *Proceedings of the Workshop on Multilingual Question Answering MLQA-2006*, 2006, Trento, Italy, pp. 1-8.

Lee, J., "Analysis of Multiple Evidence Combination," In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, Philadelphia, Pennsylvania, United States, pp. 267-276.

Lin, W. C., and H. H. Chen, "Merging Mechanisms in Multilingual Information Retrieval," In *Proceedings of the Third Workshop of the Cross-Language Evaluation Forum CLEF 2002*, 2002, Rome, Italy, pp. 175-186.

Martínez-Santiago, F., L. A. Ureña-López, and M. Martín-Valdivia, "A Merging Strategy Proposal: The 2-step Retrieval Status Value Method," *Information Retrieval*, 9(1), 2006, pp. 71-93.

Montes-y-Gómez, M., L. Villaseñor-Pineda, M. Pérez-Coutiño, J. M. Gómez-Soriano, E. Sanchis-Arnal, and P. Rosso, "A Full Data-Driven System for Multiple Language Question Answering," In *Proceedings of the 6th Workshop of the Cross-Language Evalution Forum CLEF 2005*, 2005, Vienna, Austria. pp. 420-428.

Neumann, G., and B. Sacaleanu, "Experiments on Cross-Linguality and Question-Type Driven Strategy Selection for Open-Domain QA," In *Proceedings of the 6th Workshop of the Cross-Language Evalution Forum CLEF 2005*, 2005, Vienna, Austria, pp. 429-438.

Rosso, P., D. Buscaldi, and M. Iskra, "Web-based Selection of Optimal Translations of Short Queries," *Procesamiento de Lenguaje Natural*, 38, 2007, pp.49-52.

Sangoi-Pizzato, L. A., and D. Molla-Aliod, "Extracting Exact Answers using a Meta Question Answering System," In *Proceedings of the Australasian Language Technology Workshop*, 2005, Sidney, Australia, pp. 105-112.

Savoy, J., and P. Y. Berger, "Selection and Merging Strategies for Multilingual Information Retrieval," In *Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum CLEF 2004*, 2004, Bath, UK, pp. 27-37.

Sutcliffe, R., M. Mulcahy, I. Gabbay, A. O'Gorman, K. White, and D. Slatter, "Cross-Language French-English Question Answering Using the DLT System at CLEF 2005," In *Proceedings of the 6th Workshop of the Cross-Language Evalution Forum CLEF 2005*, 2005, Vienna, Austria, pp. 502-509.

Vallin, A., B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe, "Overview of the CLEF 2005 Multilingual Question Answering Track," In *Proceedings of the 6th Workshop of the Cross-Language Evalution Forum CLEF 2005*, 2005, Vienna, Austria, pp. 307-331.