Graph-based Answer Fusion in Multilingual Question Answering

Rita M. Aceves-Pérez, Manuel Montes-y-Gómez and Luis Villaseñor-Pineda

Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y Electrónica, México.
{rmaceves, mmontesg, villasen}@inaoep.mx

Abstract. One major problem in multilingual Question Answering (QA) is the combination of answers obtained from different languages into one single ranked list. This paper proposes a new method for tackling this problem. This method is founded on a graph-based ranking approach inspired in the popular Google's PageRank algorithm. Experimental results demonstrate that the proposed method outperforms other current techniques for answer fusion, and also evidence the advantages of multilingual QA over the traditional monolingual approach.

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. In essence, a QA system is a kind of search engine that responds to natural language questions with concise and precise answers.

One major challenge that currently faces this kind of systems is the multilinguality. In a multilingual scenario, it is expected for a QA system to be able to: (*i*) answer questions formulated in various languages, and (*ii*) look for the answers in several collections in different languages.

Evidently, multilingual QA has some advantages over standard monolingual QA. In particular, it allows users to access much more information in an easier and faster way. However, it introduces additional challenges caused by the language barrier.

A multilingual QA system can be described as an ensemble of several monolingual systems [5], where each system works over a different –monolingual– document collection. Under this schema, two additional tasks are of great importance: first, the translation of questions to the target languages, and second, the combination or fusion of the extracted answers into one single ranked list.

The first problem, namely, the translation of the questions from one language to another, has been widely studied in the context of cross-language QA[1] [1, 10, 11, 14]. In contrast, the second task, i.e., the fusion of answers obtained from different languages, has only recently been addressed [2]. Nevertheless, it is important to mention that there is considerably work on combining lists of monolingual answers

* Work done under partial support of CONACYT (Project grand 43990).

[1] Cross-language QA is a special case of multilingual QA. It addresses the situation where questions are formulated in a language different from that of the (single) target collection.

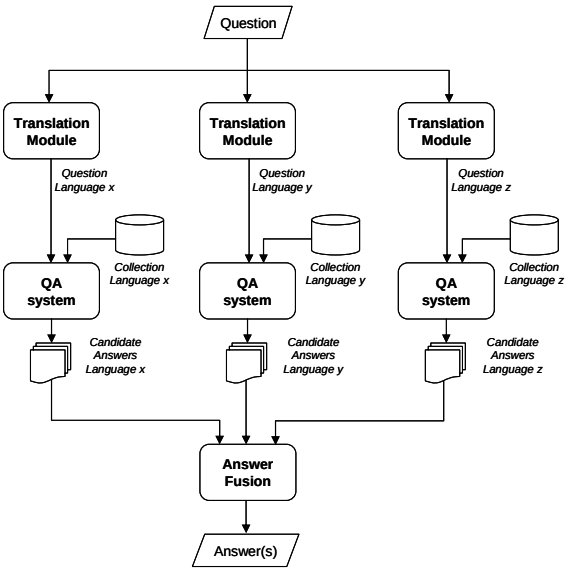

Figure 1. General architecture of a multilingual QA system

extracted by different QA systems [4, 6, 12] as well as on integrating lists of documents in several languages for cross-lingual information retrieval applications [13].

In this paper, we propose a new method for tackling the problem of answer fusion in multilingual QA. The proposed method is founded on a graph-based ranking approach inspired in the popular Google's PageRank algorithm [3]. Similar to previous related approaches, this method also centers on the idea of taking advantage of the redundancy of several answer lists. However, it models the fusion problem as a kind of recommendation task, where an answer is recommended or "voted" by similar answers occurring in different lists. In this model, the answer receiving the greatest number of votes is the one having the greatest relevance, and therefore, it is the one selected as the final answer.

The rest of the paper is organized as follows. Section 2 shows the architecture of a multilingual QA system and describes some previous works on answer fusion. Section 3 describes the proposed method. Section 4 shows some experimental results. Finally, section 5 presents our conclusions and outlines future work.

2 Related Work

Figure 1 shows a common architecture for a multilingual QA system. This architecture includes, besides the set of monolingual QA systems, a stage for question translation and a module for answer fusion.

As we previously mentioned, the problem of question translation has already been widely studied. Most current approaches rest on the idea of combining the capacities of several translation machines. They mainly consider the selection of the best in-

stance from a given set of translations [1, 11] as well as the construction of a new query reformulation by gathering terms from all of them [10, 14].

On the other hand, the problem of answer fusion in multilingual QA has only very recently been addressed by [2]. This work compares a set of traditional ranking techniques from cross-language information retrieval in the scenario of multilingual QA.

In addition, there is also some relevant related work on combining lists of monolingual answers. For instance, [6] 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 calculated in the previous step. [4] 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, [12] 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.

The method proposed in this paper is similar in spirit to [2, 4] in that it also applies a general ranking over the answers extracted from different languages, and it is comparable to [6] in that it performs an iterative evaluation process. However, our method uses a novel graph-based approach that allows taking into consideration not only the redundancy of answers in all languages but also their original ranking scores in the monolingual lists.

3 Proposed Method

The aim of the answer fusion module is to combine the answers obtained from all languages into one single ranked list. In order to do that, we propose using a graph-based ranking approach. In particular, we decide adapting the Google's PageRank algorithm[2] [3].

In short, a graph-based ranking algorithm allows deciding on the importance of a node within a graph, by taking into account global information recursively computed from the entire graph, rather than relying only on local node-specific information. In other words, this kind of ranking model put into practice the idea of voting or recommendation, where the node having the greatest number of votes is considered the most relevant one, and therefore, it is selected as the system's final output.

The application of this approach to the problem of multilingual answer fusion consists of the following steps:

1. Construct a graph representation from the set of extracted answers.
2. Iterate the graph-based ranking algorithm until convergence.
3. Sort nodes (answers) based on their final score and select the top-ranked as the system's response.

[2] It is important to mention that this algorithm has recently been used in other text processing tasks such as text summarization and word sense disambiguation [7, 8].

The following sections describe in detailed these steps. In particular, section 3.1 explains the proposed graph representation, and section 3.2 presents the graph-based ranking function.

3.1 Graph Representation

Formally, a graph $G = (V, E)$ consists of a set of nodes V and a set of edges E, where E is a subset of $V \times V$.

In our case, each node represents a different answer. This way, we will have as many nodes as answers obtained from the different languages (that is, $|V| = |A|$).

Each node $v_i \in V$ contains a set of content words $\{w_1,\ldots,w_n\}$ that describes an specific answer $a_i \in A$. In particular, we consider two levels of representation for nodes.

Direct representation: In this case, the set of content words is directly extracted from the corresponding answer. For instance, given the Spanish answer a_j = "*1 de enero de 1994*", its related node will be $v_j = \{1, \text{enero}, 1994\}$.

Extended representation: In order to make comparable the answers obtained from different languages, we extend the node representations by considering the answer's translations to all languages. For instance, if we are working with Spanish, French and Italian, then answer a_j will be represented by the node $v_j = \{1, \text{enero}, 1994, \text{janvier}, \text{gennaio}\}$.

The initial weight s_π of a node v_i is calculated in accordance with the ranking position of answer a_i in its original answer list ($r(a_i)$):

$$s_\pi{}^0(v_i) = 110 - (10 \times r(a_i)) \tag{1}$$

Using this formula, the answers at the first positions –of each language– will have a weight of 100, the second ones a weight of 90, and so on.

On the other hand, the edges of the graph establish a relation between two different answers. They mainly indicate that the answers are associated, i.e., that they share at least one content word. Obviously, the greater the number of common words between them, the greater their association value. Based on the last consideration, the weight s_σ of an edge e_{ij} between the nodes v_i and v_j is calculated as follows:

$$s_\sigma(e_{ij}) = \frac{|v_i \cap v_j|}{|v_i \cup v_j|} \tag{2}$$

where $|v_i \cap v_j|$ indicates the number of common content words of nodes v_i and v_j, and $|v_i \cup v_j|$ is the number of different words in both nodes.

Figure 2(a) shows the graph representation of the set of answers for the question "*When did the NAFTA come into effect?*". In particular, this graph includes answers in three different languages (Spanish, Italian and French), and uses the extended representation of nodes.

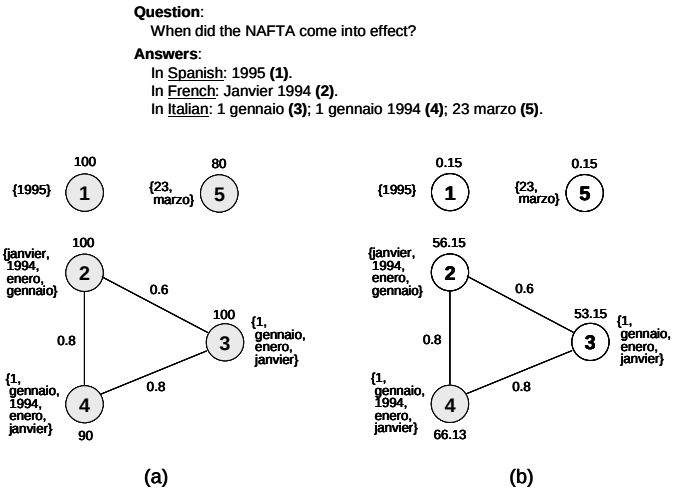

Figure 2. Example of a graph representation for answer fusion

3.2 Ranking Function

The ranking algorithm computes the scores of nodes in line with: (*i*) the number of their neighbor nodes, (*ii*) the initial weight of these neighbors (refer to formula 1), and (*iii*) the strength of their links (refer to formula 2). Therefore, the idea behind this algorithm is to reward the answers that are strongly associated to several other top-ranked responses.

Formula 3 denotes the proposed ranking function. As can be noticed, it defines the ranking algorithm as an iterative process, that –following the suggestions by Mihalcea [8]– must break off when the change in the score of one single node be less than a given specified threshold.

$$s_\pi{}^m(v_i) = (1-d) + d \times \left(\sum_{v_j \in adj(v_i)} \frac{s_\sigma(e_{ij})}{\sum_{v_k \in adj(v_j)} s_\sigma(e_{jk})} s_\pi{}^{m-1}(v_j) \right) \quad (3)$$

In this formula, $s_\pi{}^m(v_i)$ is the score of the node v_i after m iterations, $s_\sigma(e_{ij})$ is the weight of the edge between nodes v_i and v_j, and $adj(v_i)$ is a function that indicates the set of adjacent nodes to v_i.

Figure 2(b) shows the final state of the example graph after performing the ranking process. In this case, the selected answer (top-ranked node) for the question "*When did the NAFTA come into effect?*" is "*1 gennaio 1994*".

4 Experimental Evaluation

4.1 Experimental Setup

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 documents, the Italian one has 157,558, and the French one contains 129,806.

Test questions. We selected a subset of 170 factual 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 their distribution.

Table 1. Distribution of questions (by answer 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%

Monolingual QA system. We used the TOVA QA system [9]. Its selection was supported on its competence to deal with all considered languages. It obtained the best precision rate for Italian and the second best ones for Spanish and French in the CLEF-2005 evaluation exercise.

Translation Machine. For all translation combinations we used Systran[3].

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 using as a baseline the results from the best monolingual system (the Spanish system in this case). This way, it is possible to conclude about the advantages of multilingual QA over the standard monolingual approach. In addition, we also present the results corresponding to other fusion techniques[4] [2].

4.2 Results

In order to evaluate the usefulness of the proposed method, we considered the top-ten ranked answers from each monolingual QA system. Table 2 shows the results obtained when using the direct and extended graph representations. The conclusions from these results are the following.

1. Combining answers extracted from different languages sources makes possible to respond a large number of questions. In other words, multilingual QA allows improving the performance of the standard monolingual approach.
2. The proposed approach is pertinent for the task of multilingual answer fusion. In particular, using the extended representation leads to a better performance

[3] www.systranbox.com

[4] This comparison is possible because both experiments used the same set of questions as well as the same target document collections.

(14% of improvement over the baseline), since it allows better capturing the redundancy of answers across different monolingual answer lists.

Table 2. Precision achieved by the proposed graph-based approach

Method's configuration	***Precision at:***		
	1st	**3rd**	**5th**
Using the direct node representation	0.45	0.62	0.72
Using the extended node representation	**0.48**	**0.68**	**0.78**
Best Monolingual Run (baseline)	*0.45*	*0.57*	*0.64*

On the other hand, table 3 compares the results of the proposed method with those obtained by a set of traditional ranking techniques from cross-language information retrieval (for details on these techniques refer to [2]). This table indicates that the graph-based method outperforms all previously used techniques for answer fusion in multilingual QA. We believe this is mainly because our graph-based ranking approach not only takes into consideration the redundancy of the answers into the different languages, but also makes a better use of their original ranking scores in the monolingual lists.

Table 3. Comparison of several ranking techniques in multilingual QA

	Precision at:		
Method	**1st**	**3rd**	**5th**
Graph-based approach	**0.48**	**0.68**	**0.78**
RSV	0.44	0.61	0.69
RoundRobin	0.45	**0.68**	0.74
CombSum	0.42	0.66	0.75
CombMNZ	0.42	0.62	0.70

Finally, table 4 shows the evaluation results corresponding to the set of questions having their answers in more than one collection. As it was expected, the answer fusion approach had a greater impact on this subset. It is also important to notice that for this particular subset the extended node representation was much better than the direct one (10% of improvement at five positions). We consider this is because the extended representation betters capture the redundancy of answers in different languages.

Table 4. Precision on questions with answers in more than one collection

Method's configuration	***Precision at:***		
	1st	**3rd**	**5th**
Using the direct node representation	0.52	0.71	0.79
Using the extended node representation	**0.54**	**0.79**	**0.89**

5 Conclusions

This paper proposed a new method for tackling the problem of answer fusion in multilingual QA. This method is founded on a graph-based ranking algorithm that allows combining the answers obtained from different languages into one single ranked list. The algorithm takes into consideration not only the redundancy of answers but also their original ranking scores in the monolingual lists.

Experimental results showed that the proposed method is pertinent for this task. It outperforms the best monolingual performance as well as the results obtained using other current techniques for answer fusion.

As noticed from table 4, the precision at five positions is considerably greater than that for the first ranked answer. We believe this behavior is consequence of having several incorrect answer translations. In order to reduce the errors on these translations we plan to apply, as future work, some techniques for combining the capacities of several translation machines [1]. It is important to point out that the proposed scheme allows easily integrating several translations for each answer, and therefore, incrementing the possibility of retrieving the correct one.

References

1. 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.
2. Aceves-Pérez R., Montes-y-Gómez M., Villaseñor-Pineda L. Fusión de Respuestas en la Búsqueda de Respuestas Multilingüe. Procesamiento de Lenguaje Natural, Num. 38, 2007.
3. Brin S., Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30(1-7), 1998.
4. 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.
5. García-Cumbreras M. A., Ureña-López L. A., Martínez-Santiago F., Perea-Ortega J. M. BRUJA System: The University of Jaén at the Spanish Task of CLEFQA 2006. Working Notes of CLEF 2006. Alicante, España, 2006.
6. 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.
7. Mihalcea R. Graph-Based Ranking Algorithms for Sentence Extraction Applied to Text Summarization. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004). Barcelona, Spain, 2004.
8. Mihalcea R., and Tarau P. TextRank: Bringing Order into Texts. Conference on Empirical Methods in Natural Language Processing (EMNLP-2004). Barcelona, Spain, 2004.
9. 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.
10. Neumann G., Sacaleanu B. DFKI's LT-lab at the CLEF 2005 Multiple Language Question Answering Track. Working Notes CLEF 2005. Vienna, Austria, 2005.
11. Rosso P., Buscaldi D., Iskra M. Web-based Selection of Optimal Translations of Short Queries. Procesamiento de Lenguaje Natural, Num. 38, 2007.
12. Sangoi-Pizzato L. A., Molla-Aliod D. Extracting Exact Answers using a Meta Question Answering System. Australasian Language Technology Workshop. Australia, 2005.
13. Savoy J., Berger P. Y. Selection and Merging Strategies for Multilingual Information Retrieval. In Working Notes of CLEF 2004. Bath, UK, 2004.
14. 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. In Working Notes CLEF 2005. Vienna, Austria, 2005.