Enhancing Cross-Language Question Answering by Combining Multiple Question Translations

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 of state-of-the-art Cross Language Question Answering systems is the translation of user questions. This paper proposes combining the potential of multiple translation machines in order to improve the final answering precision. In particular, it presents three different methods for this purpose. The first one focuses on selecting the most fluent translation from a given set; the second one combines the passages recovered by several question translations; finally, the third one constructs a new question reformulation by merging word sequences from different translations. Experimental results demonstrated that the proposed approaches allow reducing the error rates in relation to a monolingual question answering exercise.

1 Introduction

Question Answering (QA) has become a promising research field whose aim is to provide more natural access to the information than traditional document retrieval techniques. In essence, a QA system is a kind of search engine that allows users to pose questions using natural language instead of an artificial query language, and that returns exact answers to the questions instead of a list of entire documents.

QA is a complex task that combines techniques from information retrieval, natural language processing and machine learning. Recent results from the Cross Language Evaluation Forum[1] [6] made evident this complexity showing accuracies from 68.95% (for monolingual French) to 11.5% (for monolingual Portuguese).

On the other hand, Cross Language Question Answering (CLQA) addresses the situation where the questions are formulated in a language different from that of the document collection. In this case, a user can use one language to search information from documents written in other languages. This is useful, because it would be tiresome to write the question over and over again in many languages, and also because many users have a good passive knowledge of several languages, but their active knowledge is more restricted [3].

Evidently, CLQA has many advantages over standard 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.

[1] http://clef-qa.itc.it/

Most current CLQA systems deal with the language barrier problem by translating the questions to the document's language [4, 9, 10, 12, 13]. This solution is very intuitive and seems effective, but it is too sensitive to the translation errors. This effect was noticeable in the QA report from the last CLEF edition [6]. There, the results corresponding to the best system were 67.89% of accuracy for the French monolingual task and 45.26% for the English-French bilingual exercise [6]. These results indicate that the translation errors caused a relative drop in accuracy of about 33%.

Given the great impact of the translation errors in the final answer accuracy, recent CLQA systems apply various techniques in order to reduce the error rates of the translation module. For instance, [5] performs a triangulated translation using English as a pivot language, and [13] translates the question keywords using a bilingual dictionary as well as EuroWordNet. Some other works combine the capacities of several translation machines[2]. In particular, [12] generates a term-by-term translation combining two different translation machines and a dictionary, and [9] constructs an expanded "bag of words" query gathering terms from several question translations as well as their synonyms extracted from EuroWordNet.

In this paper, we propose some new methods to tackle the language barrier problem in CLQA. Similar to previous approaches, these methods also center around the idea of combining the capacities of several translation machines. However, they consider not only the construction of a new query reformulation by gathering terms from several translations, but also the selection of the best translation from a given set and the combination of passages recovered by different question translations. Furthermore, the proposed methods have a great potential to be used in many CLQA scenarios since they do not make use of additional language-dependent resources such dictionaries or ontologies.

The rest of the paper is organized as follows. Section 2 describes the three proposed methods for tackling the language barrier problem in a CLQA application. Section 3 presents the evaluation results. Finally, section 4 gives our conclusions and describes some future work.

2 Proposed Methods

As we mentioned, one major problem in current CLQA systems is the translation of the user questions. In order to reduce the drop in accuracy caused by the translation mistakes, we propose to combine the capacities of multiple translation machines. This idea is mainly supported in the following assumptions:

1. Given that machine translation is a complex task, there is still not available a perfect translation machine.
2. Different translation machines tend to produce –slightly– different and –partially– correct question translations.

[2] Similar ideas have been proved in other fields. For instance, [11] proposes a method that combines several WSD systems by selecting the one best for each specific word.

3. The more frequent a term is in the set of translations, the more chances that the original word has been translated correctly.

Based on these assumptions we designed three different methods (or architectures) for CLQA. The first method selects the most fluent translation from a given set, and then delivers it to a monolingual QA system. The second method combines the passages recovered by several question translations in one single set, and then uses these passages to extract the answer to the given question. Finally, the third method constructs a new question reformulation by merging word sequences from different translations, and then sends this new query to a monolingual QA system.

The following subsections describe in detail the proposed methods.

2.1 Method 1: "Selecting the Best Translation"

Figure 1 shows the general scheme of this method. It consists of three basic steps. First, the question is translated to the target language (i.e., the language of the document collection) using a number of translation machines. Second, all translations are evaluated and the best one is selected. Finally, the selected translation is given to a monolingual QA system in order to obtain the desired answer.

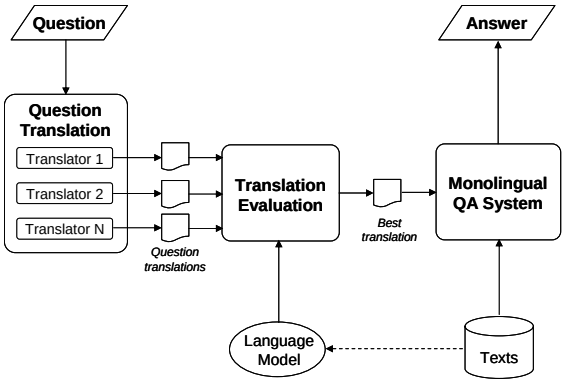

Figure 1. CLQA Method considering the selection of the best translation

An accepted criterion to evaluate the quality of translations indicates that the most fluent output text corresponds to the best translation. A known mechanism to determine the fluency of a given translation is to measure it pertinence to a predefined language model [1]. The language model judges the probability that a test data –in this case a translation– fits to that language. In our particular case, we propose to measure the pertinence of the translations with respect to the target document collection.

2.1.1 Translation evaluation

The pertinence of a translation to the target document collection is based on how much it fits in the collection *n*-gram model. In order to quantify this attribute we

apply a general *n*-gram test on the translation. An *n*-gram test computes the entropy (or perplexity) of some test data –the question translation– given an *n*-gram model. It is an assessment on how probable is to generate the test data from the *n*-gram model[3]. The entropy is calculated as follows:

$$H = -\frac{1}{Q}\sum_{i=1}^{Q}\log P(w_i|w_{i-1}, w_{i-2},..., w_{i-N+1})$$

where w_i is a word in the *n*-gram sequence, $P(w_i|\ w_{i-1},\ w_{i-2},\ldots,w_{i-N+1})$ indicates the probability of observing w_i right after the occurrence of the *n*-gram $w_{i-1},\ w_{i-2},\ldots,w_{i-N+1}$, Q is the number of words of the test data, and N is the order of the *n*-gram model.

The final score for a translation is expressed by its perplexity, defined as $B = 2^H$. In this case, the lowest perplexity value indicates the most probable expression on the target collection, and therefore, the most pertinent translation.

2.2 Method 2: "Combining Passages from Several Translations"

In order to take advantage of all translations we consider the combination of passages recovered by all of them. Figure 2 shows the general scheme of this method. It considers the following procedures. First, the user question is translated to the target language by several translation machines. Then, each translation is used to retrieve a set of relevant passages. After that, the retrieved passages are combined in order to form one single set of relevant passages. Finally, the selected passages are analyzed and a final question answer is extracted.

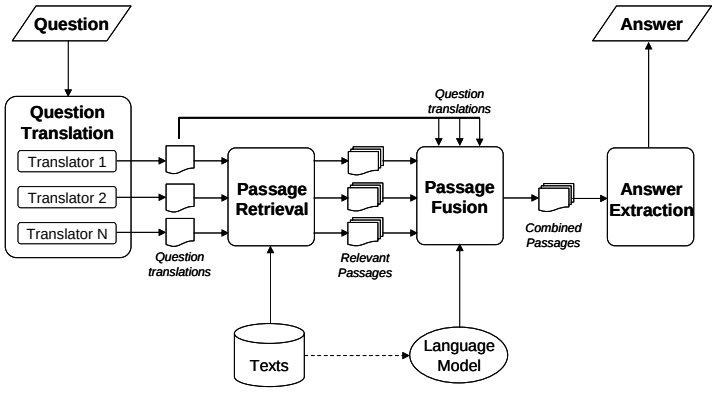

Figure 2. CLQA Method considering the combination of passages

The main step of this method is the combination of the passages. This combination is based on the pertinence of the translations to the target document collection. This pertinence, as in the previous method, expresses how a given translation fits in

[3] The *n*-gram model was constructed using the method described in [15].

the *n*-gram model calculated on the target document collection. The idea is to combine the passages favoring those retrieved by the more pertinent translations.

2.2.1 Passage Combination

This module combines the retrieved passages from each translation in one single set. Its purpose is to favor passages recovered by the more pertinent translations. The following formula is used to calculate the number of passages from a given translation that will be included in the combined passage set.

$$E_x = \frac{k}{\sum_{i=1}^{n} \frac{1}{B_i} \times B_x}$$

In this formula E_x indicates the number of selected passages from the translator x, that is, the extension of x in the combined set. B_x is the perplexity of the translator x (refer to section 2.1.1), n is the number of translation machines used in the experiment, and k indicates the number of passages retrieved by each translator as well as the total extension of the combined set[4].

2.3 Method 3: "Constructing a Question Reformulation"

After analyzing several question translations we could notice that (*i*) correct word sequences tend to occur in more than one translation, i.e, they are repeated, and that (*ii*) slightly different translations may contain different correct translations for the same word, i.e, they tend to use some synonyms. Based on these observations we propose to combine several question translations in one single question reformulation. This reformulation contains all words occurring in more than one question translation.

Figure 3 shows the general scheme of this method. It considers three basic steps. First, the user question is translated to the target language by several translation machines. Then, all translations are combined to form a new single question reformulation. Finally, this question reformulation is given to a monolingual QA system in order to obtain the desired answer. The following subsection describes the procedure to combine a set of question translations.

2.3.1 Combining translations

The combination of translations aims to capture the common words among the different translations and to maintain in some way the relative order of the words in the question reformulation. This idea is different than other previous methods [8, 12] in that it goes beyond the bag-of-words approach, since it considers word sequences as well as it frequency of occurrence.

The procedure to combine the translations is as follows: Given a set of question translations *T*:

---

[4] In the experiments we set k = 20, which corresponds to the best performance rate of our monolingual QA system [7].

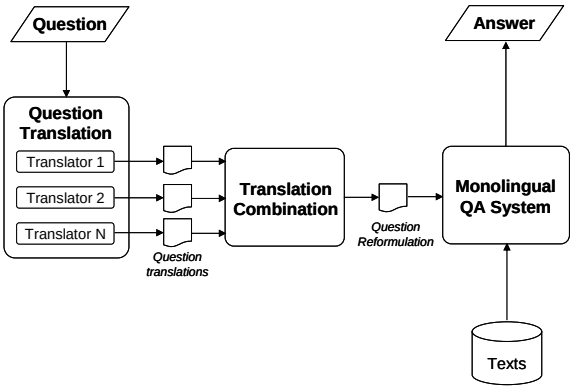

Figure 3. CLQA Method using a question reformulation

1. *Extract the set of maximal frequent word sequences from T*. A maximal frequent word sequence is a sequence of words that occurs more than a predefined threshold and that is not a subsequence of another frequent sequence.
2. *Select the more frequent sequence as the initial query reformulation*.
3. *Add to the initial query reformulation the content words from other sequences*. These words must not be contained in the initial query reformulation.

3 Experimental Results

3.1 Experimental Setup

For the experimental evaluation we used a set of 286 factoid questions extracted from the CLEF Multi-Eight corpus as well as the CLEF Spanish document collection consisting of 454,045 news documents.

The evaluation considered three bilingual experiments: English-Spanish, French-Spanish and Italian-Spanish. For translating the questions to Spanish we used three different translation machines[5]: Systran, Worldlingo, Fretranslation.

For the experiments we used the passage retrieval and answer extraction components of the TOVA question answering system [7]. We selected this system because it was one of the best in the Spanish QA task at the 2005 edition of the CLEF. We also used the data-mining tool described in [2] in order to compute the maximal frequent word sequences required by one of the methods. In this case, we established a threshold $\sigma = 2$, which indicated that a word sequence was frequent if it was contained in at least two different translations.

[5] www.systranbox.com, www.worldlingo.com, www.freetranslation.com

3.2 Results

As we previously mentioned, one major problem of state-of-the-art Cross Language Question Answering systems is the translation of user questions. Several QA reports [6, 14] indicate that the translation errors cause an important drop in accuracy for cross-language tasks with respect to the monolingual exercises. Based on this fact, we evaluated the impact of our methods by measuring the fall of accuracy[6] in the answer extraction caused by the question translation in relation to the Spanish monolingual QA task.

Table 1 shows the fall of accuracy, indicated as an error rate, corresponding to the three bilingual experiments. In this table, the first three columns indicate some baselines, which correspond to the error rates generated by each translation machine when they were used independently. On the other hand, the last three columns show the error rates obtained when we applied each one of the proposed methods.

Table 1. Error rates with respect to the Spanish monolingual task

	Baselines *(a single translation machine)*			*Our Methods*		
	TM1	TM2	TM3	Best Translation	Passages Combination	Query Reformulation
English-Spanish	25%	28%	27%	14%	12%	10%
French- Spanish	28%	30%	28%	17%	16%	15%
Italian- Spanish	30%	45%	41%	41%	24%	13%

The results indicate that our three methods reduced –in the majority of the cases– the fall in accuracy, and produced lower error rates than using one single translation machine. For instance, for the English-Spanish exercise we could reduce the error rate from 25% (corresponding to the best single translation machine) to just 10% using the query reformulation method. For the French-Spanish task, the error rate moved from 28% to 15%, while for the Italian-Spanish we reduced it from 30% to 13%.

It also is important to notice that the worst results correspond to the Italian-Spanish exercise. We believe these results were consequence of the bad quality of the used translators (with error rates from 30-45%). In particular, this situation greatly affects the performance of the best translation method, since no translation fit well to the language model.

Finally, it is also important to point out that the best methods were those that combine the capacities of all translations. Specifically, the query reformulation method produced the best results. We consider this performance is due because it simulates a kind of query expansion, retaining just the most confident words of all translations.

[6] The accuracy indicates the percentage of correctly answered questions. It is calculated as the radio between the number of found answers and the number of questions.

4 Conclusions and Future Work

In this paper we presented three different methods to tackle the language barrier problem in CLQA. These methods consider the selection of the most fluent translation form a given set, the combination of the passages recovered by different question translations, and the construction of a question reformulation by merging word sequences from several translations.

The experiments indicated that the three proposed methods allowed reducing the fall in accuracy, and produced lower error rates than using any translation machine independently. They also gave some evidence about that the best methods were those that combine the capacities of all question translations, namely, the passage combination method and the query reformulation approach. These results confirmed our hypothesis that all translations are partially correct and that using information from all of them allows identifying answers that could not be find using one single question translation. Nevertheless, it is important to emphasize that our conclusions are not completely general, since our results are in some extent dependent to the used QA system, especially to the passage retrieval system, as well as to the target language and the document collection.

As future work we plan to do some additional experiments in order to determine some parameters of the methods. In particular, we plan to: (*i*) use more translation machines in order to determine the number and the quality of the selected translators; (*ii*) experiment with other target languages; and (*iii*) evaluate the performance of the proposed methods when using some other QA systems.

Acknowledgements. This work was done under partial support of CONACYT (Project Grant 43990). We also like to thanks to the CLEF organizing committee as well as to the EFE agency for the resources provided.

5 References

1. Callison-Burch C., and Flournoy R. A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines. In Proceedings of the Machine Translation Summit VIII, Santiago de Compostela, Spain, 2001.
2. García-Hernández Rene A., Martínez Trinidad José Francisco, Carrasco-Ochoa Jesús Ariel: A Fast Algorithm to Find All the Maximal Frequent Sequences in a Text. CIARP 2004: 478-486.
3. Gonzalo J. Scenarios for interactive cross-language retrieval systems. Proceedings of the Workshop of Cross-Language Information Retrieval: A Research Roadmap Workshop held at the 25th Annual International ACM SIGIR Conference. Tampere, Finland, 2002.
4. Jijkoun Valentin, Mishne Gilad, Rijke Maarten de, Schlobach Stfan, Ahn David, Muller Karin. The University of Amsterdam at QA@CLEF 2004. In CLEF, editor, Proceedings CLEF-2004 Lecture Notes in Computer Science, pp. 321-324, 2004.
5. Laurent Dominique, Séguela Patrick, and Nègre Sophie. Cross lingual question answering using QRISTAL for CLEF 2005. In Working Notes, CLEF Cross-Language Evaluation Forum, Vienna, Austria. 2005.
6. Magnini B., Giampiccolo Danilo, Forner Pamela, Ayache Christelle, Osenova Petya, Peñas Anselmo, Jijkoun Valentin, Sacaleanu Bogdan, Rocha Paulo, Sutcliffe Richard.

Overview of the CLEF 2005 Multilingual Question Answering Track. CLEF 2006, Alicante, España, 2006.
7. 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. CLEF 2005, Vienna, Austria, 2005.
8. Neumann Günter and Sacaleanu Bogdan. DFKI's LT-lab at the CLEF 2005 multiple language question answering track. In Working Notes, CLEF Cross-Language Evaluation Forum, Vienna, Austria. 2005.
9. Pablo-Sánchez César de, González-Ledesma Ana, Martínez-Fernández José Luis, Guirao José Maria, Martinez Paloma, and Moreno Antonio. MIRACLE's 2005 approach to cross-lingual question answering. In Working Notes, CLEF Cross- Language Evaluation Forum, Vienna, Austria. 2005.
10. Perret L., "Question answering system for the French language", In CLEF, editor, Proceedings CLEF-2004 Lecture Notes in Computer Science, pp. 295-303.
11. Saarikoski H., Legrand S., Gelbukh A. Defining Classifier Regions for WSD Ensembles Using Word Space Features. MICAI-2006. Lecture Notes in Artificial Intelligence, N 4139, Springer, 2006.
12. Sutcliffe Richard F. E., Mulcahy Michael, Gabbay Igal, O'Gorman Aoife, White Kieran, Slatter Darina: "Cross-Language French-English Question Answering using the DLT System at CLEF 2005" In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien.
13. Tanev Hristo, Negri Matteo, Magnini Bernardo, and Kouylekov Milen. The DIOGENE question answering system at CLEF-2004. In Working Notes, CLEF Cross-Language Evaluation Forum, pages 325–333, Bath UK. 2004.
14. 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. CLEF 2005, Vienna, Austria, 2005.
15. Villaseñor-Pineda L., Montes-y-Gómez M., Pérez-Coutiño M. and Vaufreydaz D.. A Corpus Balancing Method for Language Model Construction. Conference on Intelligent Text Processing and Computational Linguistics CICLing-2003, D.F., Mexico, February, 2003. Lecture Notes in Computer Science, vol. 2588, Springer, 2003.