Cross-language Question Answering: The Key Role of Translation

S. Larosa[1], J. Peñarrubia[2], P. Rosso[3], M. Montes-y-Gomez[4]
[1]Dipartimento di Informatica e Scienze dell'informazione
Università degli Studi di Genova, Italy
2000s036@educ.disi.unige.it
[2]Facultad de Informática, Universidad Politécnica Valencia, Spain
jlpenarr@upvnet.upv.es
[3]Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es
[4]Laboratorio de Tecnologias de Lenguaje
Instituto Nacional de Astrofisica, Óptica y Electrónica, Mexico
mmontesg@inaoep.mx

Abstract

The goal of a Question Answering (QA) system is to provide inexperienced users with a flexible access to the information allowing them for writing a query in natural language and obtaining a concise answer. Cross-language QA systems allow the user for querying in a language different than the language in which documents are written. In this paper, we illustrate a case study to understand how much the translation of the questions may reduce the accuracy of a QA system. The main goal is to investigate whether more machine translators could be used in order not to rely just on one translation and to choose the best one on a statistical basis.

1. Introduction

Nowadays, the Web has become our main information repository: nearly all kind of information (digital libraries, newspapers collections, etc.) is available in electronic format. These documents may satisfy almost every information need. Therefore, rather than Question Answering (QA) systems which are based on sophisticated linguistic analyses of both questions and candidate answers, it makes sense to use a language-independent approach, which is supported by the data redundancy of the Web [1]. The main idea is that questions and answers are commonly expressed using the same words, and that the probability of finding a simple (lexical) matching between them increases with the redundancy of the Web [2, 3, 4].

In recent years, the combination of the Web growth and the explosive demand for better information access has motivated the interest in developing QA systems. Many are the efforts made both by academic institutions as well as well known research companies like IBM, which recently developed the prototype of the Piquant (Practical Intelligent Question Answering Technology) search engine [5].

Documents on the web are written in more than 1,500 languages. Therefore, it is useful to provide an inexperienced user with a flexible access to the information allowing for writing a question in her mother tongue, and obtaining a concise answer [6].

In this paper, we illustrate a study for a Cross-Language Question Answering in which the questions are made in a certain language whereas the documents are written in a different one. In order to tackle the problem of the translation of the questions, a combination of translators should be used. The paper is structured as follow. Section 2 describes the Cross-language Web-based QA system and the experiments we carried out. Section 3 illustrates the language-independent approach we have been investigating and the section 4 shows some preliminary results. Finally, some conclusions are drawn in the section 5.

2. Cross-Language Web-Based QA System

The system we used was developed at the Language Technologies laboratory of the INAOE at Mexico [7]. Given a question, the QA system makes combinations of its words, searching for these new queries on the Web through a search engine's browser (e.g. Google). For each of the new query reformulations (obtained manipulating the order of the words of the question), the system collects a certain number of snippets (the

snippet is the part of a relevant document that the browser retrieves which contains almost all the words of the query). Finally, possible answers are extracted on a statistical basis, and a final ranking of candidates is returned. Therefore, the main steps of the QA system are: query reformulation (verb movement, bag of words, components [7]), snippets recollection, and answer extraction. In case of Cross-language QA, a translation preprocess is needed in order to translate the questions from the source language into the target language of the documents. In order to extract the most frequent n-grams (sequences of n words) from the snippets (each n-gram is defined as a possible answer to the given question), we used a statistical criterion which ranks them by decreasing likelihood of being the correct answer. The method which is used for the n-gram extraction and ranking is based on regular expressions. A compensation factor is applied in order to avoid favoring short n-grams with respect to large ones. The method extracts the twenty most frequent unigrams which satisfy a given typographic criteria (i.e., words starting with an uppercase letter, numbers and names of months), determines all the n-grams (from bigrams to pentagrams, built from the set frequent unigrams), ranks the n-grams based on their compensated relative frequency, and finally selects the top five n-grams (candidates as possible answers).

The compensated relative frequency of a *n*-gram $g(n) = (w_1 \ldots w_n)$ is computed as follows [7]:

$$P_{g(n)} = \sum_{i=1}^{n} \sum_{j=1}^{n-i} \frac{f_{j(i)}}{\sum_{\forall x \in G_i} f_{x(i)}}$$

where G_i is the set of *n*-grams of size *i*, $|G_i|$ indicates the cardinality of this set, $j(i)$ is the *n*-gram *j* of size *i* contained in $g(n)$, and $f_{j(i)}$ is the frequency of occurrence of this *n*-gram. The QA system has been tested in monolingual (Spanish, Portuguese and Italian) [7,8] as well as in Cross-language (Catalan-Spanish and Arabic-English) tasks [9]. For the Catalan-Spanish and Arabic-English QA Cross-language experiments, the original corpus of the Cross-Language Evaluation Forum (CLEF)-2003 [10] (mainly focused on answering factual queries, i.e., those having a simple named entity as the answer) was manually translated into Catalan and Arabic. Thereafter, the translation of the questions was made using the SALT Valencian-Spanish translator [11] and the TARJIM Arabic-English translator [12], respectively.

The precision of correct answers obtained with the questions translated from Catalan into Spanish was approximately half of that obtained directly with the Spanish questions. It has to be mentioned that both languages have many similar words, and in some cases even searching on the Web with the question in Catalan, the retrieved snippet was in Spanish.

In the Arabic-English Cross-language experiments, we compared the results obtained querying the QA system with the original corpus in English and with that one obtained automatically after the Arabic-English translation. In Table 1 it is possible to appreciate that the number of questions correctly answered decreased of more than one third (in the best case of the verb movement reformulation). The table gives an idea of how much the accuracy of the results may decrease due to the translation process of the questions.

Table 1. Precision and MRR measures

Questions	Bag words	Comp.	Comp no 1st word	Comp no 1st and 2nd words	Verb mov.
English (original)	17.1% 0.12	24.4% 0.19	26.7% 0.20	22.0% 0.16	**39.5%** **0.31**
English (from Arabic)	6.0% 0.04	2.4% 0.02	7.4% 0.06	8.4% 0.06	**10.7%** **0.08**

The Mean Reciprocal Rank (*MRR*) measure was also used to fully evaluate the performance of the system:

$$MRR = \frac{1}{n} \sum_{i=1}^{n} r_i$$

The MRR measure takes into account what is the ranking of the extracted answer (the contribution of a question, which is not obtained an answer for, is 0): n is the total number of test questions and ri is the reciprocal of the rank (position in the answer list) of the first correct answer. For instance, if the correct answer is in the second position, ri = 0.5, whereas if it is in the third then ri = 0.33. In the case the correct answer does not occur in the list of the top five n-grams, then ri =0.

At the moment of writing this paper, some other Cross-language experiments have been carrying out (Urdu-English, Persian-English, and Italian-Spanish) in order to study how much the translation pre-process of the questions may decrease the performance of the QA system for other language combinations. No matter how much exactly the accuracy decreases in each Cross-language task: it is no doubt that the translation has a key role in the final performance of the system. Therefore, the way to improve the quality

of the translation of the questions needs to be investigated. In the next section a first statistical attempt is described.

3. Combining Translations

A very important step for a Cross-language QA system is the translation of a question from a language source to a destination one. Generally, majority of QA systems use online translators, but the quality of their translations is often not very good and this has a negative impact on the QA system efficiency. We suggest an approach which uses more than one translator and selects the best translation. Two methods were implemented: *Word-Count* and *Double Translation*. Word-Count exploits the redundancy of terms in all the translations, and the translation with the highest number words in common (in other words the most similar) will be chosen. To establish the number of common words and calculate the similarity among the translations, two formulae have been chosen: the *Dice* and the *Cosine* formulae. With Word-Count and the Dice formula we make an intersection of the translations to find the number of common words.

In order to illustrate the two language-independent approaches, we describe them using the following examples of translated question from Italian into Spanish with four different translators [13]:

"Che cosa significa la sigla CEE?"
("What does the acronym EEC mean?")

1. ¿Qué significa la sigla CEE?
2. ¿Qué cosa significa siglas el EEC?
3. ¿Qué significa la CEE de la abreviación?
4. ¿Qué cosa significa la pone la sigla CEE?

Therefore, the Dice formula is used to establish the degree of similarity among the translations in order to rank them:

$$sim(t_i, t_j) = \frac{2 \times len(t_i \cap t_j)}{len(t_i) + len(t_j)}$$

where:
- t_i and t_j are the two different translations;
- $len(t_i \cap t_j)$ indicates the number of common words of both translations;
- $len(t_k)$ represents the number of words of translation t_k.

To get a corresponding similarity value for every translation, the similarity between a translation and the others has to be calculated using the previous formula (the partial results will be added together in order to obtain its similarity value). For instance, to get the similarity of the first translation we do: $sim(t_1,t_2)$+ $sim(t_1,t_3)$ + $sim(t_1,t_4)$. The translation with the highest value is chosen. To increase the accuracy of the choice of the best translation, n-grams are used (an *n*-gram is a sequence of *n* words). If for instance there are two translations which have the same identical words but with a different order, n-grams allows for calculating their similarity values. Examples of 2-grams of the sentence below are:

"Qué significa la sigla CEE?"
("What does the acronym EEC mean?")

"Qué significa" "significa la" "la sigla" "sigla CEE"

The *Word-Count method* was implemented also using the cosine formula to calculate the similarity degree. In this model, translations are represented as vectors in a *t*-dimensional space (*t* is the general number of index terms or keywords). The keywords weights are calculated using a scheme-like Term Frequency – Inverse Document Frequency (tf-idf) [14]. Examples of translated question with four different translators are:

"Qual' è la capitale della Repubblica del Sud Africa?"
("What is the capital of the Republic of South Africa?")

1. ¿Cuál es la capital de la República de la Sur África?
2. ¿Cuál es entendido ellos de la república de la África del sur?
3. ¿Cuál es la capital de la República del Sur una Africa?
4. ¿Cuál es el capital de la república del sur Africa?

The list of keywords is: "cuál", "es", "la", "capital", "de", "república", "sur", "África", "entendido", "ellos", "del", "una", "africa","el"

We get the list of keywords of all translations (in order to define the dimensionality of the vector space), and then measure the weight of every keyword for every translation using the following formula:

$$t_{ij} = f_{ij} \times \log(1 + \frac{n_i}{N})$$

where:
- t_{ij} indicates the weight of word *i* at translation *j*;
- f_{ij} is the normalized frequency of word *i* in the translation *j;*
- *N* is the total number of translations;

- n_i is the number of translations containing the word *i*.

Once the vectors have been found, the next step is the calculation of the similarity degree among all the translations by using the following formula:

$$sim(t_i, t_j) = \frac{\left(\sum_{\forall k} t_{ik} \times t_{jk}\right)}{\sqrt{\sum_{\forall k} t_{ik}^{\ 2}} \times \sqrt{\sum_{\forall k} t_{jk}^{\ 2}}}$$

In the formula t_{ik} and t_{jk} represent two generic vector weights. The translation with the highest value is chosen. The final calculation is done as follows:

Translation1 = $sim(t_1,t_2) + sim(t_1,t_3) + sim(t_1,t_4)$
Translation2 = $sim(t_2,t_1) + sim(t_2,t_3) + sim(t_2,t_4)$
Translation3 = $sim(t_3,t_1) + sim(t_3,t_2) + Sim(t_3,t_4)$
Translation4 = $sim(t_4,t_1) + sim(t_4,t_2) + Sim(t_4,t_3)$

With the *Double Translation method*, every question in Italian is translated into Spanish and then retranslated back into Italian. Four translators are used and the translation whose results are more similar to the original question will be chosen. The *Dice* and the *Cosine* formulae are used in this case as well. The algorithms used are those previously illustrated. Example of original question and double translations are:

"Che cosa significa la sigla CEE?"
("What does the abbreviation EEC mean?")

1. ¿Che cosa significa la sigla CEE?
2. ¿Che cosa significa le abbreviazioni il EEC?
3. ¿Che significa il CEE dell'abbreviazione?
4. ¿Che cosa ha importanza la mette la sigla di CEE?

As we already mentioned, the methods are totally statistical, and therefore language-independent. At the moment of writing this paper, the application of the methods to other pairs of language other than Italian-Spanish is under investigation (e.g. Catalan-Spanish and Arabic-English [9]). The only limitation to these methods derives from the availability of translators in the source language.

4. Experiments

In the experiments we carried out, we translated 450 factual question derived from the CLEF 2003 competition. Four different translators were used (only two of these allow a direct translation from Italian to Spanish). The following tables show the percentage of success and the number of question which were properly translated in every experiment.

Table 2. Word-count, Dice formula

1-Gram	*2-Grams*	*3-Grams*
51.33%	51.11%	51.55%
231/450	230/450	232/450

Table 3. Double-Translation, Dice formula

1-Gram	*2-Grams*	*3-Grams*
46.66%	49.11%	50.22%
210/450	221/450	226/450

Table 4. Word-count, Cosine formula

1-Gram	*2-Grams*	*3-Grams*
48.66%	49.33%	50.00%
219/450	222/450	225/450

Table 5. Double-Translation, *Cosine* formula

1-Gram	*2-Grams*	*3-Grams*
45.77%	48.44%	49.11%
206/450	218/450	221/450

From these experiments we have observed that some translators made bad translations (in particular those that not allow a direct translation from the source language into the target one). The machine translator which obtained the best results is *PowerTranslationPro* (55.33%). This baseline was better than our best results (51.55%) which were obtained with the Word-Count method. Nevertheless, the preliminary results we obtained seem to be promising. In fact, an optimal combination among the *Word-count* and *Double Translation* methods could increase the percentage of success. We estimate that it should be possible to obtain approximately an increase of up to 20% of the system's performance. This is due to the fact that the choices obtained from two methods are not the same. Finally, we carried out another experiment in order to investigate how to combine the methods. In this last experiment we make a comparison between the methods and the baseline. The questions were separated into the following categories: Date, Person, Organisation, Location, and Measure.

The table 6 shows the best results obtained by the methods, in comparison with the baseline machine translator (PowerTranslationPro). For every method appear only the best percentage among the methods. The numbers in bold means that a method was capable to reach a better performance then a baseline. For the Person category, our approach obtains the same results of the baseline, whereas for the Organisation and the Measure categories, the percentage of the correctly translated questions is higher. Probably, with the help of these results, we can make a good combination

between *Word-Count* and *Double Translation* and improve the percentage of success.

Table 6. Questions separated for categories

	Date	*Person*	*Organization*	*Location*	*Measure*
Number of Questions	44	71	26	61	77
WordCount Dice and 1-gram	--	--	**46%**	59%	**58%**
WordCount Dice and 2-gram	--	--	--	--	**58%**
Double Trans Dice and 2-gram	61%	--	--	--	--
Double Trans Dice and 3-gram	61%	**64%**	--	--	--
Double Trans Cosine and 3-gram	61%	--	--	--	--
Baseline	70%	64%	42%	72%	40%

5. Conclusions

In this paper we investigated the possibility of improving the question translation preprocess of a Cross-language QA system. Two totally statistical and language-independent methods were described. The preliminary results seem to be promising an for some of the studied categories were better than those obtained by the baseline. Further experiments are needed to find an optimal combination among the methods and, therefore, increase the percentage of success. As further work, it would be also interesting to use the JIRS passage retrieval system [15] in order to fully take advantage of the redundancy of the Web during the validation of the translations.

Acknowledgments

The work was partially supported by the R2D2 (CICYT TIC2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054) research projects and CONACYT 43990.

References

[1] E. Brill, J. Lin, M. Banko, and S. Dumais, "Data-intensive question answering", Proc. TREC-10, 2001.

[2] E. Hovy, L. Gerber, U. Hermajakob, M. Junk, and C. Lin, "Question answering in Webclopedia", Proc. TREC-9, 2000.

[3] C. Kwok, O. Etzioni, and D. Weld, "Scaling question answering to the Web", Proc. of the WWW Conference, 2001.

[4] J. Lin, J., "The Web as a resource for question answering: perspectives and challenges", Proc. of the 3rd Int. Conf. on Language Resources and Evaluation (LREC), 2002.

[5] IBM Piquant Question Answering system, at: http://www.research.ibm.com/compsci/spotlight/nlp/

[6] J. Vicedo, "Los Sistemas de Búsqueda de Respuestas desde una Perspectiva Actual", Revista Iberoamericana de Inteligencia Artificial, 2004.

[7] M. Del Castillo, M. Montes y Gómez, and L. Villaseñor, "QA on the web: A preliminary study for Spanish language", Proc. of the 5th Mexican Int. Conf. on Computer Science (ENC), Colima, Mexico, 2004.

[8] L. Villaseñor-Pineda, M. Montes-y-Gómez and A. del Castillo, "Búsqueda de respuestas basada en redundancia: un estudio para el Español y el Portugués", Proc. Taller de Herramientas y Recursos Lingüísticos para el Español y el Portugués, IX Ibero-American Conf. on Artificial Intelligence IBERAMIA 2004, Puebla, Mexico, November, 2004.

[9] P. Rosso, A. Lyhyaoui, J. Peñarrubia, M. Montes y Gómez , Y. Benajiba, and N. Raissouni, "Arabic-English Question Answering", Proc. of Information Communication Technologies Int. Symposium (ICTIS), Tetuan, Morocco, June 2005.

[10] Cross-Language Evaluation Forum (CLEF) European consortium: http://www.clef-campaign.org

[11] SALT Valencian-Spanish Translator, available at: http:// www.cult.gva.es/salt/salt_programes_salt2.htm

[12] TARJIM Arabic-English Translator, available at: http://tarjim.ajeeb.com/ajeeb/default.asp

[13] S. Larosa, M. Montes y Gómez, P. Rosso and S. Rovetta, "Best Translation for an Italian-Spanish Question Answering System",Proc. Of Information Communication Technologies Int. Symposium (ICTIS), Tetuan, Morocco, June 2005.

[14] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.

[15] J. Gómez, M. Montes y Gómez, E. Sanchis and P.Rosso, "A Passage Retrieval System for Multilingual Question Answering Answering", LNCS, Springer Verlag, TSD Int. Conf, Brno, Check Republic, September 2005 (accepted; to be published).