# Use of Free On-line Machine Translation for Interactive Cross-Language Question Answering

Angel Zazo, Carlos G. Figuerola, José Luis A. Berrocal, and Viviana Fernández Marcial

REINA Research Group – Universidad de Salamanca
C/ Francisco de Vitoria 6-16, 37008 Salamanca, SPAIN
http://reina.usal.es

**Abstract.** Free on-line machine translation systems are employed more and more by Internet users. In this paper we have explored the use of these systems for Cross-Language Question Answering, in two aspects: in the formulation of queries and in the presentation of information. Two topic-document language pairs were used, Spanish-English and Spanish-French. For each of these, two groups of users were created, depending on the level of reading skills in document language. When machine translation of the queries was used directly in the search, the number of correct answers was quite high. Users only corrected 8% of the translations proposed. As regards the possibility of using machine translation to translate into Spanish the text passages shown to the user, we expected the search of the users with little knowledge of the target language to improve notably, but we found that this possibility was of little help in finding the correct answers for the questions posed in the experiment.

## 1 Introduction

Question Answering (QA) is one of the most advanced facets in information retrieval. It searches for precise answers to specific questions. The idea is to find a minimum text fragment that will answer the question, using an extensive document collection to do so. When the question and the documents are in different languages, this is called Cross-Language Question Answering (CL-QA). The process becomes complicated if documents are in a language in which the user is rather unskilled.

On-line machine translation systems are free tools becoming more well-known and used by Internet users. For the Cross-Language Evaluation Forum (CLEF) 2005 interactive track (iCLEF), we explored the use of machine translation (MT) for interactive CL-QA. Our intention was to reproduce the normal situation of users with little knowledge of the language of the documents, unable to form the query correctly or to correctly understand a possible answer. In many cases these users resort to on-line MT services to satisfy their information needs. We carried out the experiment with two language pairs: Spanish-English and Spanish-French, in order to see the dependence of the results on the target language. We focused on two aspects:

1. *The formulation and refinement of the queries.* We wished to analyse the behaviour of users employing an interactive CL-QA system when they can initiate or refine the searches using their own language or the language of the documents.
2. *The possibility of using MT to translate the information shown to the user.* We wished to observe the behaviour of users with little knowledge of the language of the documents when they have the possibility of translating them to their own language and whether this possibility improves the accuracy of the system.

To have a suitable basis for comparison the experiments were carried out with two groups of users for each language pair, each with a different reading level in the language of the documents. This goal was to be able to analyse the behaviour of both types of users.

## 2 The CLIR System

We actually used as a cross-language information retrieval (CLIR) system a standard document retrieval system that made monolingual searches in the language of the documents. It was based on the vector space model, with different adaptations to translate the questions to the language of the documents and the documents to the language of the user. The system was the same one we used in iCLEF2004 [1], with some modifications.

Text passages were used instead of complete documents, the same as last year, but the possibility of seeing the context of the passage, i.e. the complete document, was excluded, although as we know [1] this reduces the accuracy of the system. This year the passages were made up of at least 50 words (including stop words). If a paragraph had fewer words, it was joined to the following one, and so on as necessary to complete a passage of at least 50 words.

Last year the CLIR system also used an on-line MT system to translate to the target language the questions written in Spanish, but refinement of the searches was only permitted by means of a very limited mechanism of term suggestion (which, by the way, was not greatly appreciated by the users). In this year's experiments the users could refine their searches with greater freedom, both in Spanish and the target language. Interaction with the user was done using standard web forms. The interface permitted the refinement of the searches and the examination of the passages retrieved, with the possibility of translating the latter to the language of the user.

## 3 The Experiment

We followed the iCLEF guidelines [2] for carrying out the experiment, which indicated what the queries were, how the search should be carried out, which document collections could be used, the questionnaires and the time limit. The

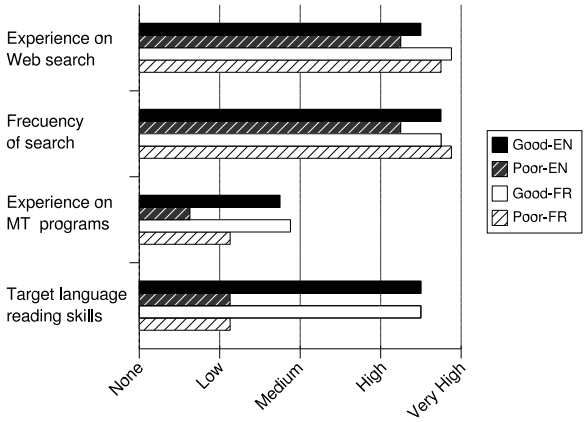


**Fig. 1.** First questionnaire prior to the experiment.

Spanish version of the queries was used in the experiments, and the corresponding document collections in English and French supplied by the CLEF organization.

The experiment was carried out with four groups of users. For each pair of query-document languages, Spanish-English and Spanish-French, two groups of users were created, each with a different reading level in the language of the documents: good and bad (or *poor*). The users were university students whose mother-tongue was Spanish. The groups were named Good-EN, Poor-EN, Good-FR and Poor-FR. The groups with the prefix "Good" were made up of students from the degree course in Translation at the University of Salamanca. These users actually carried out monolingual searches, and their tests served as a referent point for those of the "Poor" group.

Figure 1 shows results of initial questionnaire previously to the searches. For all groups a great deal of experience in Web search was reported. For all users, the frequency of search was close to once or twice a day. Experience of MT programs was small for all groups, but smaller for "Poor" groups. Notice the difference in the knowledge of the target language of both types of groups.

### 3.1 Machine Translation

On-line machine translation systems are free tools being handled more and more by Internet users. In our experiment we used two of these systems to translate queries or terms from Spanish into the target language and passages from English/French into Spanish:

- Spanish–English: *Google Linguistic Tools* (`http://translate.google.com`)
- Spanish–French: *Systran Online* (`http://w3.systranbox.com`)

Initially we used Google because it did not impose length limit on the input text, but it did not have the Spanish-French translation pair. We thus used

Systran for that pair, although sometimes the connection with the Systran server stalled. To avoid this type of effect, in the experiment we did not compute the time employed in the translation process (generally between 1 or 2 seconds for both systems, except when there were problems with the Systran connection).

### 3.2 Reference and Contrastive Systems

The reference system (*System A*) was a standard document retrieval system based on the vector space model performing monolingual searches in the language of the documents. To formulate the query the users could write the question in Spanish or directly in the target language (Fig. 2). The button labelled "Traducir_y_Buscar" (*Translate and Search*) translated the text entered in the first field into the target language and then immediately made the search. The button labelled "Buscar" (*Search*) carried out the search directly using the terms entered into the second entry field. Before using the system, the users were told how it worked: it was a simple term driven system, and thus the users could use terms instead of questions or sentences. It must be pointed out that, in order to facilitate the typing of the question, in each initial search the field corresponding to Spanish was filled in automatically with the text of the query, and the users were free to use it as such, change it if they wished or enter their own terms in the target language.

Once the search had been carried out, the user was shown an ordered list of retrieved passages (Fig. 3). Within the 5 minutes time limit established for each search, the users could refine the search using the entry fields in the upper part of the interface, both in Spanish and the target language. The lower part of the interface contained fields to fill in the answer and the degree of confidence. The users could abandon the search at any time ('nil' answer), by clicking onto the checkbox button labelled "No encuentro la respuesta" (*I cannot find the answer*). After 5 minutes, a window appeared showing only the lower part of the interface, permitting the user a final chance to write the answer.

The contrastive system (*System B*) was identical to the reference system, except that it allowed for the possibility of translating the passages into Spanish. The button "Traducir este pasaje" (*Translate this passage*) only appeared in this system (see Fig. 3). When clicking this button the original passage and its translation were shown.

## 4 Results and Discussion

### 4.1 Difference between Target Languages

Figure 4 shows the strict accuracy and the average searching time for each group. A priori we did not expect much difference between the groups with a good knowledge of the target language, but it turned out to be large, both in strict accuracy and in average searching time. The users of the Good-FR group needed much less effort to find correct answers to the questions. We believe that the

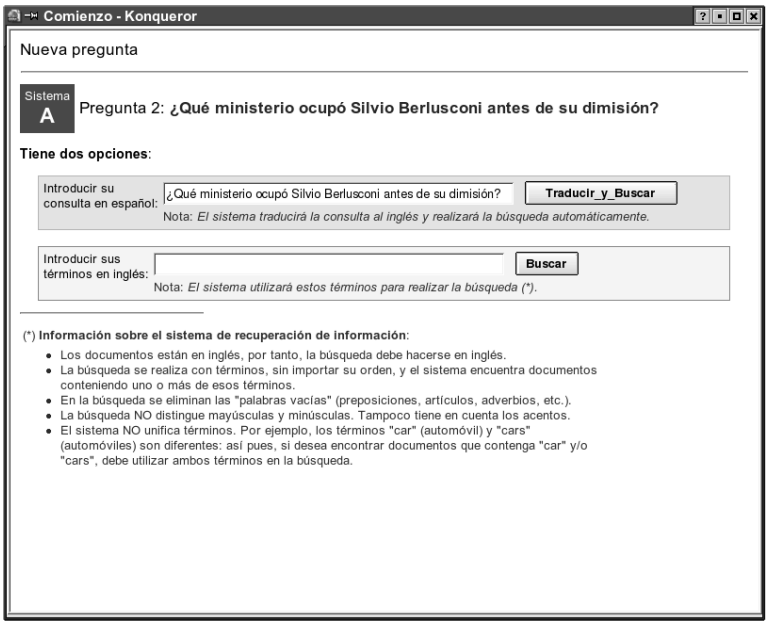


**Fig. 2.** Initial search of a question.

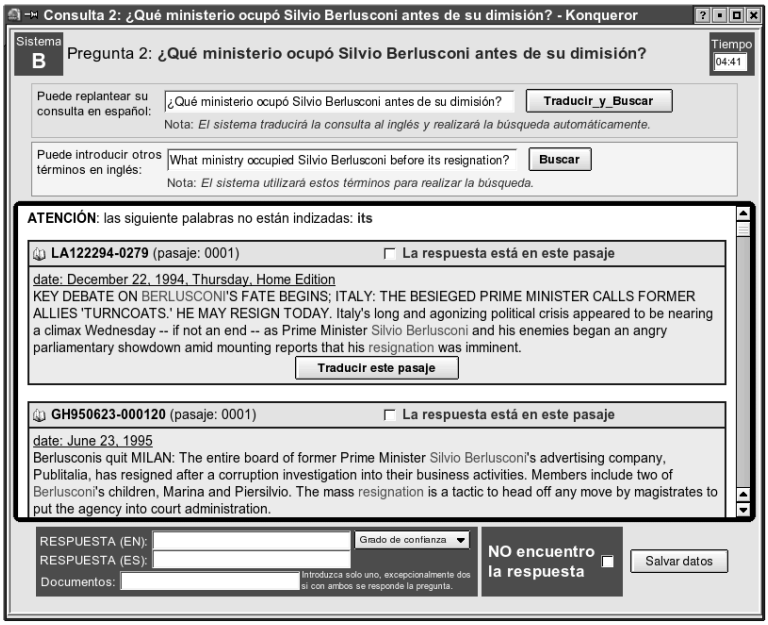


**Fig. 3.** Ranked list of passages showed to user. Refinement is possible in both reference and contrastive systems.

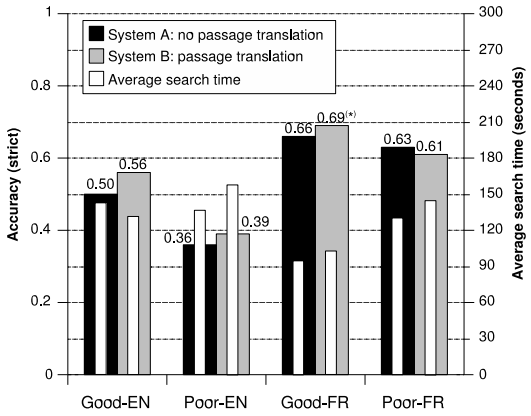


**Fig. 4.** Strict accuracy and average search time. (*) One 'nil' answer was mistakenly assessed as correct in contrastive system for Good-FR group.

reason for this is the division into passages of text. Considering the set of 16 questions of the experiment, the division in text passages had better results for the document collection in French than for the one in English. We should recall that in the experiment the possibility of seeing the context of the passages, i.e. the complete document, was intentionally excluded. If the context of the passages had been available, we believe the difference between the two systems would have been less.

In accordance with what we expected, the groups with a better reading level in the target language obtained greater accuracy than the "Poor" groups, although for both French groups the difference was very small. This was because French is closer to Spanish than English is: Spanish users with little knowledge of French and English can better understand a possible answer in a text written in French than in one written in English. The number of passages translated with the contrastive system by the "Poor" groups was greater for the English group (see Table 1), which corroborates the above affirmation.

### 4.2 Difference between Reference and Contrastive Systems

An analysis of the strict accuracy of each group showed that there was no significant difference between the reference system and the contrastive system (Fig. 4). Neither was the search time very different. We had expected the "Poor" groups to have greater accuracy with the contrastive system, but there was hardly any difference. Curiously enough, the difference between the systems was greater for the "Good" groups, although they hardly used the possibility of the translation of passages of the contrastive system. The average number of passages per question translated into Spanish with the contrastive system can be seen in Table 1. No user of the Good-FR group used the possibility of translating the passages into Spanish. Only one user of the Good-EN group had several passages

**Table 1.** Average number of translated passages per topic.

| Good-EN | Poor-EN | Good-FR | Poor-FR |
|---|---|---|---|
| 0.13 | 3.56 | 0 | 1.11 |

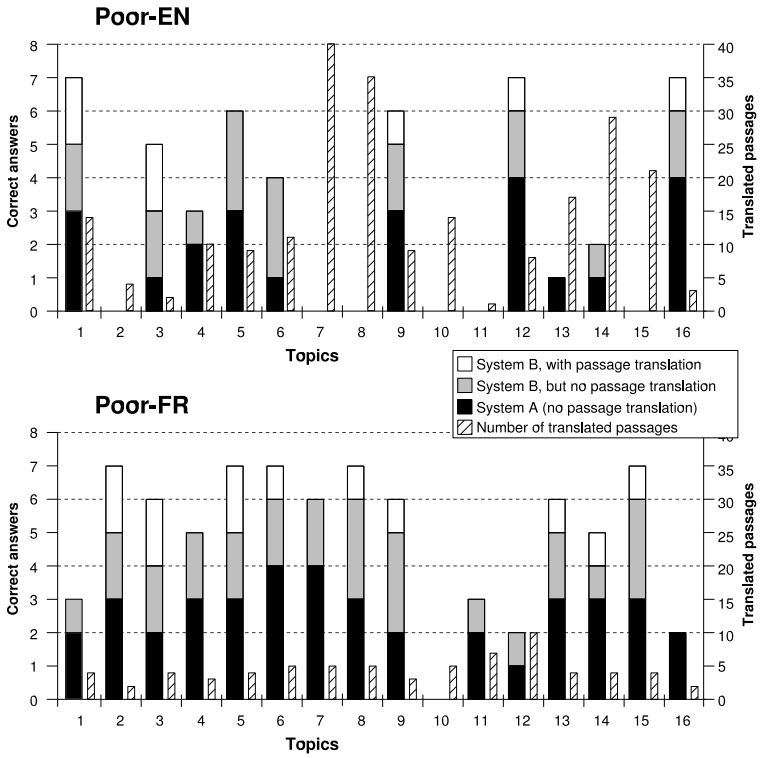


**Fig. 5.** Correct assessed answers and number of translated passages per topic.

translated from English into Spanish, but said he did it to see the quality of the translation, not because he needed help in the search for answers.

For the "Poor" groups, Figure 5 shows the number of correct answers with the reference and contrastive systems for all the questions of the experiment. In the contrastive system the correct answers obtained were differentiated after the user employed the option of translating the passages. Figure 5 also includes the total number of passages translated for each question in the contrastive system. We can see that for both target languages the number of correct answers obtained after having carried out the translation of the passages was low in comparison with the total: 7 out of 48 for English (14.58%) and 12 out of 79 for French (15.19%). It seems that the translation of the passages was of little help to the users in finding correct answers for the particular set of questions in this experiment.

The post-system questionnaires for the reference and contrastive systems were very similar. In general, the contrastive system obtained a better assessment, and the difference was larger in the "Poor" groups. However, these groups did not obtain a significant advantage in accuracy, and the accuracy was less even for the Poor-FR group. According to the results of the final post-search questionnaire, all groups found both systems easy to understand and use. As to which of the systems was considered better in general, the "Good" groups found no differences between them, since, except for one user, they did not make use of the translation possibility of the contrastive system. However, the "Poor" groups indicated that they thought the contrastive system was much better.

All users remarked that the possibility of translating the passages was highly appreciated. Also they noted that it was very useful for locating the possible answer the fact that search terms were displayed in different colour than text.

### 4.3 Query Formulation and Refinement

**Initial Search.** Before beginning the search, the users had to choose how to formulate the question: in Spanish or in the language of the documents. As was to be expected, the "Poor" groups began their searches almost exclusively in Spanish. Seventy percent of the users of the Good-EN group and 71% of the Good-FR group also began their searches in Spanish. The percentage is very similar for all the questions. Nevertheless, there were differences among the users: in general, each user employed almost exclusively one of the two methods to begin all their searches. Several of the users who began their searches in Spanish indicated that it was very convenient to use the "Traducir_y_Buscar" button, since the question appeared automatically in the associated field: they let the system make the translation into the target language and then they changed it if they considered it incorrect or if they did not find suitable answers.

**Refinement.** If the users did not find answers with the initial search, they could refine the search. The refining could be done in Spanish or in the target language. Most of the users refined their searches in the target language. We were able to observe that the way refining is carried out greatly depends on each user. For example, user number 3 in the Poor-EN group carried out most of the refinements in that group. Several users of the Good-EN and Good-FR groups did not refine their searches, although several of their answers were 'nil' answers.

**Quality of the Translation.** For a term driven document retrieval system, the syntactic or grammatical quality of a translation is of little importance. What is really important is that the translation of the terms be correct in their context. When the search was initiated in Spanish, the number of correct answers without the need to refine the search was high: 71 out of 116 for English (61.21%), and 117 out of 165 for French (70.91%). In the experiment, corrections of the machine translations were made in only a few cases: 21 times in a total of 251 translations (initial search + refinement) from Spanish to English (8.37%) and

21 times in a total of 256 translations from Spanish to French (8.20%). As was to be expected, the groups with a good level in the target language made more corrections than the "Poor" groups. In their comments, several users in the Good-FR group pointed out that they were pleasantly surprised by the quality of the translations of the questions into French.

The errors in translation depended on the language pair in each test. In the use of Google for the translation of Spanish to English, the difficulties were found mainly with three terms: "Economía" (*Economy*), which Google translated as "Economi'a"; the term "Turquesa" (*Turquoise*) which it did not translate because it had a capital letter; and the term "Universo" (*Universe*) which it translated as "Universal". For the Spanish to French translations with Systran, the difficulties were also found with three terms: "Turquesa", which Systran translated as "Turque" (*Turkish*) instead of "Turquoise", and the terms "Noruega" (*Norway*) and "Eduardo", which were not translated because they contained capital letters (the correct translations would have been "Norvège" and "Edouard").

### 4.4 Failure Analysis

There were fewer answers judged as correct for the tests with English as the target language. We believe this was due to the worse results in the division of passages for the document collection in English and the impossibility of seeing the complete document. If the context of the passages had been available, the accuracy of the systems would have been greater. This justifies in part the high number of unsupported answers in the groups with English as the target language (11%). Other aspects to be taken into account are the incorrect translations, affecting 3 questions, and also imprecise answers for some of the questions. When French was the target languages the errors were due mainly to incorrect translations.

### 4.5 'Nil' Answers

The mean time for the 'nil' answers was quite uniform for the four groups: about 4 minutes, i.e. the users gave up their searches before the fixed maximum time ran out. This was more pronounced in the second half of each test and denotes tiredness with the experiment. Several users indicated that the test was somewhat tedious: there were many questions, some of them long and complicated.

### 4.6 Topic 9

The answers to the question "*¿Con el nombre de qué enfermedad se corresponde el acrónimo BSE?*" were judged differently by the English and French assessors. For our tests with the Spanish-English language pair only two answers were affected (one for each group): the accuracy was barely affected by this.

## 5 Conclusions

The use of free on-line machine translation for interactive CL-QA was explored in two important aspects: in the search process and in the visualization of information. In the first it was found that with direct use the machine translated questions the number of correct answers was high. The fundamental difference was found in language pairs: the results were better for Spanish into French than for Spanish into English. In our case, this differences lies in the fact that French is much closer to Spanish than English is, and the quality of the translation is higher between French and Spanish. The quality of the translation depends on the original and target languages. Other authors have also had this result [3, 4].

Regarding visualization, we expected that users with little knowledge of the document language whould to obtain higher accuracy since they could use MT to translate the passages shown to them in another language into Spanish. In fact, all the users manifested in their comments that they valued very positively the possibility of translating into Spanish the passages in the contrastive system. The users in the "Poor" groups thought that the contrastive system was far better. However, for both target languages the number of correct answers obtained after having the passages translated was low in comparison with those obtained without using this option. It seems that the possibility of translating the passages was of little help in finding the correct answers for the questions of the experiment.

Finally, the differences obtained with the languages of the experiments must also be pointed out. The strict accuracy for the Good-EN and the Good-FR groups was quite different. In our view this was due to the importance of the division of the text passages when it is not possible to see the context of the information. If the context of the passages had been available, the accuracy of the systems would have been better in general, and the difference between these groups would have been less.

## References


1. Figuerola, C.G., Zazo, A., Alonso Berrocal, J.L., Rodríguez, E.: Interactive and bilingual question answering using term suggestion and passage retrieval. CLEF 2004, Lecture Notes in Computer Science **3491** (2005) 363–370
2. Gonzalo, J., Clough, P., Vallin, A.: Overview of the CLEF 2005 interactive track (In this volume)
3. Jones, G.J., Lam-Adesina, A.M.: Exeter at CLEF 2001: Experiments with machine translation for bilingual retrieval. CLEF 2001, Lecture Notes in Computer Science **2406** (2002) 59–77
4. Kraaij, W.: TNO at CLEF-2001: Comparing translation resources. CLEF 2001, Lecture Notes in Computer Science **2406** (2002) 78–93