iCLEF2005 at REINA-USAL: Use of Free On-line Machine Translation Programs for Interactive Cross-Language Question Answering

Ángel F. Zazo Rodríguez, Carlos G. Figuerola, José Luis Alonso Berrocal and Viviana Fernández Marcial
Grupo de Recuperación de Información Avanzada (REINA)
Universidad de Salamanca (USAL)
37008 Salamanca - SPAIN
reina@usal.es

Abstract

For the iCLEF experiment of CLEF 2005 at University of Salamanca, we have explored the use of free on-line machine translation (MT) programs for the interactive Cross-Language Question Answering process (CL-QA), in two aspects: query formulation and displaying information. Two question-document language pair have been used: Spanish-English and Spanish-French. We used the GOOGLE Linguistic Tools at `http://translate.google.com` for the translation between first language pair, and SYSTRANSBox at `http://w3.systranbox.com` for the second. Our cross-language information retrieval system is a standard document retrieval system performing monolingual searches. Passages has been used rather complete documents, but the possibility of examining the context of a passage is intentionally excluded, although it reduces accuracy. For each question-document pair two groups of users were constituted depending if they have good or poor reading skills in document language. The use of on-line MT program for the translation of Spanish queries into document language obtains a high number of right answers for all groups, but better for Spanish-French pair groups, due to French is more closed language to Spanish than English, so the quality of the translations between Spanish and French is better. For our interactive CL-QA experiment, few corrections of translations were necessary in general, fewer for French target language.

The use of MT to translate passages from English/French into Spanish in the displaying information process seems not improve the accuracy of the system. We expected that poor reading skills groups should obtain very much accuracy using this possibility, but difference is scarce when they don't use it. In general, all users reported that the possibility of translation the passages is highly appreciated. More experiment must be carried out to achieve other conclusions.

1 Introduction

For our participation in the interactive CLEF track we have explored the use of free on-line machine translation (MT) for the interactive Cross-Language Question Answering (CL-QA) process. On-line MT systems are within reach of any Internet user, and are, in fact, more and more known and used. We want to reproduce the common situation when users have poor reading skills in the language of documents, and cannot formulate a correct query or understand correctly a possible answer. In many cases these users use on-line MT systems to satisfy their information needs. We have made the experiment for two question-document language pair: Spanish-English and Spanish-French, in order to see if great differences exist.

Our experiment has focused on two important aspects:

1. In the formulation and refinement of queries. We want to see the behaviour of the users who use an interactive CL-QA system when several possibilities are offered to them:

- to input or refine queries in a well-known language (Spanish) and use MT to translate the queries into document language,
- to input or refine queries in the language of documents,
- or any combination of the previous ones.

2. In the possibility of using free on-line MT systems to translate the information shown to the user. We want to see the behaviour of the users with poor skills in the language of documents when they have the possibility of using a free on-line MT system to translate the retrieved information into its own language, and if this possibility improves the accuracy of the system. Machine translation of the information showed to users can only be used in the contrastive search system.

In order to have a suitable comparison base it is necessary to make the experiments with two groups of users for each pair of languages: (i) users with a good reading skills in the language of questions and documents, and (ii) users with poor reading skills in the language of documents. The reason has been to analyse so much the behaviour of both types of users, like the dependency of language pairs in this experiment.

2 The CLIR system

Our CLIR system is a standard document retrieval system performing monolingual searches (in document language), not a QA system. It is based on the vectorial space model, with some adaptations to translate the questions (into document language) and documents (into user language) using a free on-line machine translation program. The base CLIR system is the same that we have used in CLEF2004. Some modifications have been made to carry out the experiments this year:

- Passages has been used rather complete documents, just like last year, but the possibility of examining the context of a passage is intentionally excluded, although it reduces accuracy [1]. The division on passages was different this year: it has been made dividing the documents in paragraphs more than 50 words (including stop words). If a paragraph has fewer words, the passage was formed joining the next one. Usually the last passage of a document has less than 50 words. The averaged passage length was 71.5 words for English collection, and 85.6 words for French collection.
- Last year the CLIR system automatically translate the questions using a machine translation program: no query reformulation was allowed, except for a very limited term suggestion mechanism (it was low appreciated by users). This year we have interest in the formulation and refinement of search questions when a machine translation program is possible to use.
- The interaction with the system was carried out through using Web pages with forms. Much iteration can be done to reformulate the question, examine retrieved passages, and the possibility to translate passages into the language of the user.

3 The Experiment

Our experiments follows the iCLEF 2005 experiment design[1], which prescribes how to conduct the search, what are the questions, document collections, questionnaires and time limitations.

3.1 Tests data

We have used the Spanish version of the question set. The experiment has been realized for two question-document language pair: one with the document collection in English, and the other with the French one. The collections comprise news data from 1994 and 1995 taken from the Los Angeles Times and the Glasgow Herald for English language, and the Schweizerische Depeschen Agentur and Le Monde for French.

[1] iCLEF website: `http://nlp.uned.es/iCLEF`

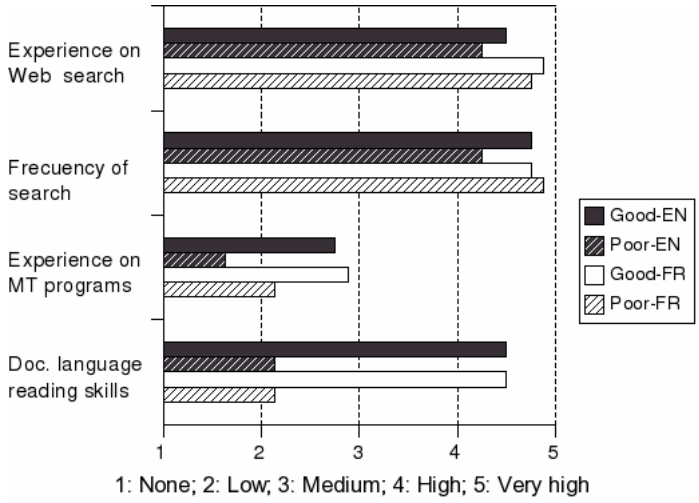

Figure 1: Initial questionnaire.

3.2 Users

Four groups of eight users have made the experiment. For each question-document language pair (Spanish-English and Spanish-French) was formed two groups with different reading skills in the language of document: good and poor. All users were native speakers of Spanish, mainly university students. Groups were denoted as Good-EN, Poor-EN, Good-FR and Poor-FR. Groups with 'Good' prefix was formed mainly with students of Translation Science at University of Salamanca. Really, they made monolingual interactive information retrieval, but their tests were reference tests for 'Poor' groups.

Figure 1 shows results of initial questionnaire previously to the searches. For all groups a great deal of experience in Web search was reported. All users reported an average of about 6 years in on-line searching experience. The frequency of search was closed to one or twice searches at day for all groups. Experience in using MT programs is low for all groups but less for groups with poor reading skills in the language of documents. Notice that the reading skills in the language of documents are very different for 'Good' and 'Poor' groups.

3.3 Machine Translation

On-line machine translation programs are free tools increasingly used for Internet users. In our experiment, two on-line programs have been used to translate questions/terms from Spanish into document language, and passages from English/French into user language:

- Spanish⇔English: GOOGLE Linguistic Tools at `http://translate.google.com`.
- Spanish⇔French: SYSTRANS Online at `http://w3.systranbox.com`.

At first, we only wanted to use GOOGLE because it sets no length limit to the input text, but no Spanish-French language pair it has. We used SYSTRANS for that pair, but sometimes the Internet connection with SYSTRANS was stalled. To avoid time effects in the experiment we not computed the time for translation process (usually near 1 or 2 seconds for both systems, except for stalled connections).

3.4 Reference and contrastive systems

The reference system (*Sistema A*) is a standard document retrieval system based on the vectorial space model performing monolingual searches in document language. To formulate the question (see Figure 2), users can type queries in Spanish or in document language. The "*Traducir_y_Buscar*" (translate&search) labelled button translates the Spanish query/text written in first field into the language of documents and immediately performs the search. The "*Buscar*" (search) labelled button directly performs the search using the terms written in the second field. Users were instructed on how the system works first to use it: our system is a simple term driven system, so they could use terms instead of questions or sentences.

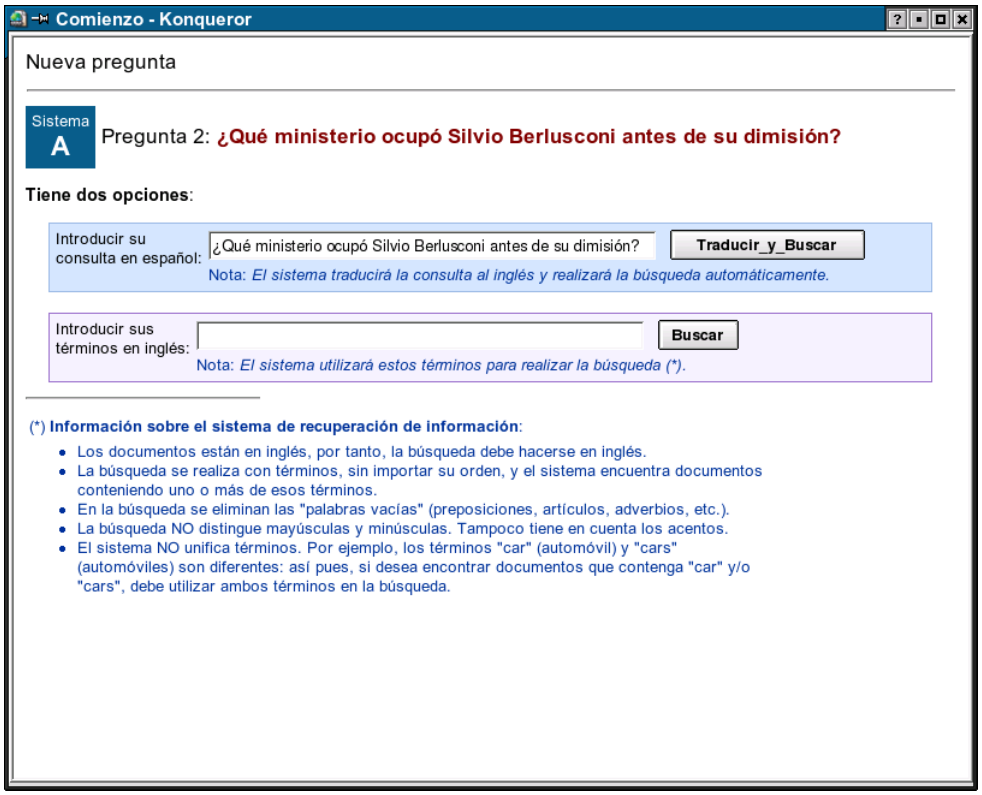

Figure 2: Initial search process for a topic.

Once searching finished, a ranked list of passages is showed in a frame of the interface, every one displays the document identification, passage number and date of document, and a checkbox button to mark the passage if it contains the answer. Remark that showed passages are in document language. Users can reformulate the query at the upper part of the interface, either in Spanish or in document language at any time within 5 minutes limit. The lower part contains fields to fill the answer and a selector for the confidence. Users can abandon a question search at any time ('nil' answer), clicking the checkbox with the label "*No encuentro la respuesta*" (I don't find the answer). When time expired, the system shows a last chance window to write the answer (only shows the lower part of the interface).

The contrastive system (*Sistema B*) is identical to reference system, except for the possibility of translating passages into Spanish. The button "*Traducir este pasaje*" (translate this passage) only appears in this system (see Figure 3); when clicking it the frame shows original and translated passage (see Figure 4).

4 Results

Table 1 shows exact and lenient accuracy, averaged searching time and averaged number of passages translated into Spanish in contrastive system.

5 Comments

5.1 Differences between reference and contrastive systems

For all groups no significant difference exists in exact accuracy and averaged searching time between reference and contrastive systems (see Figure 5). We expected that 'Poor' groups should obtain very much accuracy with contrastive system, but difference is scarce. Perhaps it is due to these groups were constituted with heterogeneous users. We discuss users effects later. Curiously, for 'Good' groups the

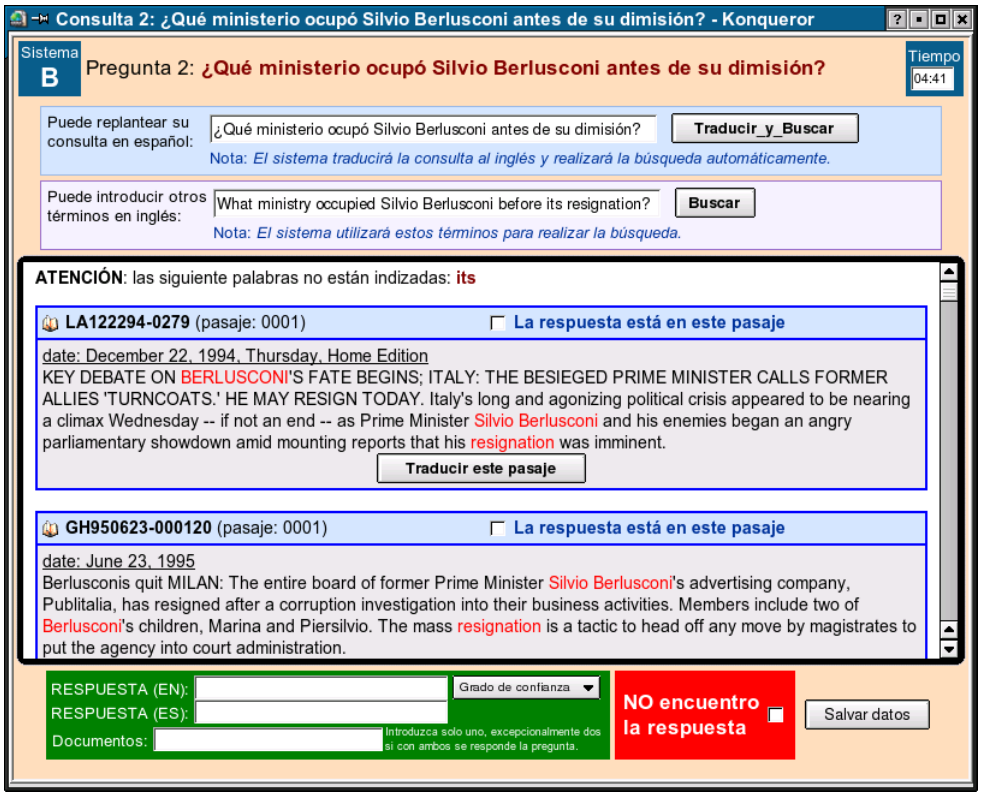

Figure 3: Ranked list of passages showed to user. Refinement is possible in both reference and contrastive system.

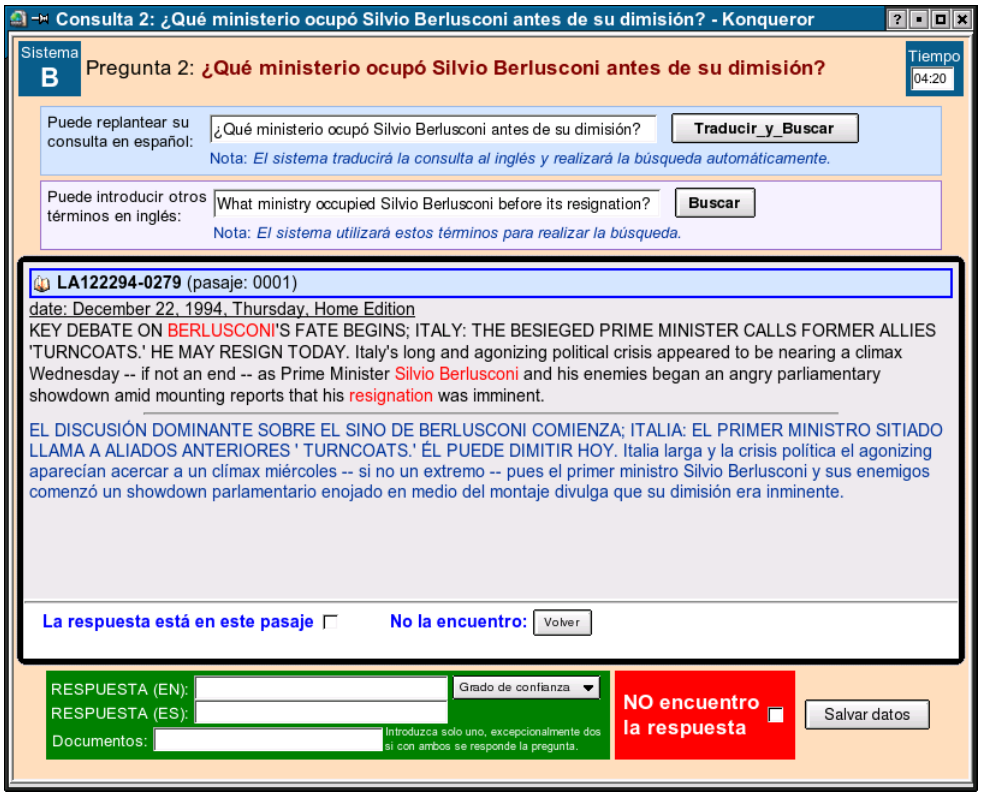

Figure 4: Contrastive system: translating passages into Spanish.

Group	System	Accuracy		Time	# Translated
		Exact	Lenient	(av.)	passages (av.)
Good-EN	Passages	.50	.53	142.8	—
	Translated	.56	.56	131.5	0.13
Poor-EN	Passages	.36	.42	136.6	—
	Translated	.39	.45	157.7	3.56
Good-FR	Passages	.66	.67	94.6	—
	Translated	.69(*)	.73	102.8	0.00
Poor-FR	Passages	.63	.70	130.1	—
	Translated	.61	.66	144.6	1.11

(*) One 'nil' answer was mistakenly assessed as correct.

Table 1: Results of the experiment.

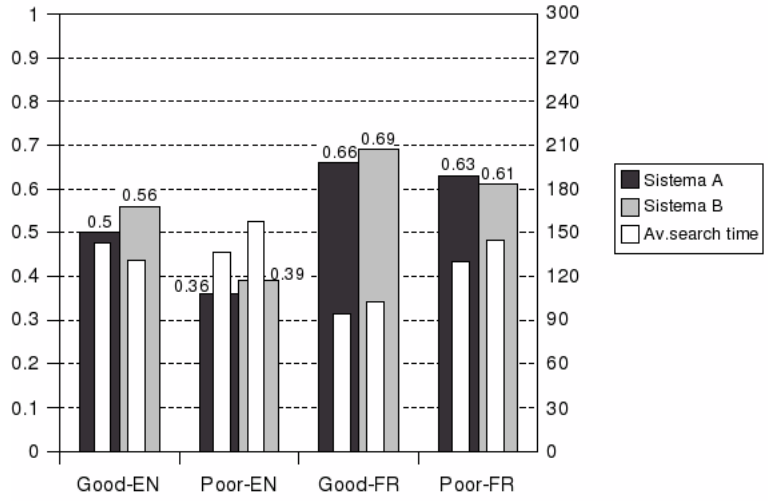

Figure 5: Differences between reference and contrastive systems.

difference in exact accuracy between systems is bigger, still yet they have been hardly performed passage translations with contrastive system.

As was to be expected, groups with good reading skills in document language achieve better accuracy, although for French the difference is very small respect to 'Poor' group. It is because of French is more closed language to Spanish than English was (Spanish users with low skills in French and English can understand better a possible answer in a French text than in English one). The averaged number of translated passages per question in contrastive system is bigger in English tests than in French ones.

For all groups the post-system questionnaires for both reference and contrastive systems are very similar (see Figure 6). In general, contrastive system obtains better appreciation, which is stronger for 'Poor' groups. However, these groups no obtain a significant improvement in accuracy with that system (even for the Poor-FR group the accuracy is worse).

For all groups the final post-search questionnaires indicate that both systems are easy to learn and use. 'Good' groups also do not find differences in which system is better overall (they hardly use the possibility of translation a passage). However, 'Poor' groups marked contrastive system was far the best. In general, all users remarked that the possibility of translation the passages is highly appreciated. Also they noted that was very useful for localizing the possible answer that search terms were displayed in different colour than text.

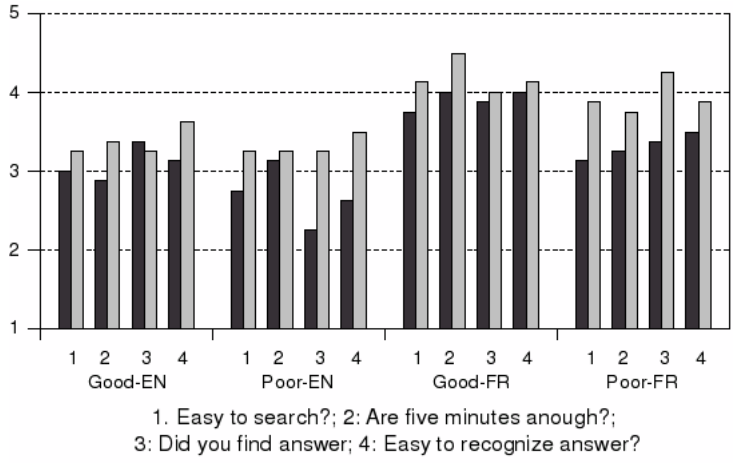

Figure 6: Post-system questionnaires.

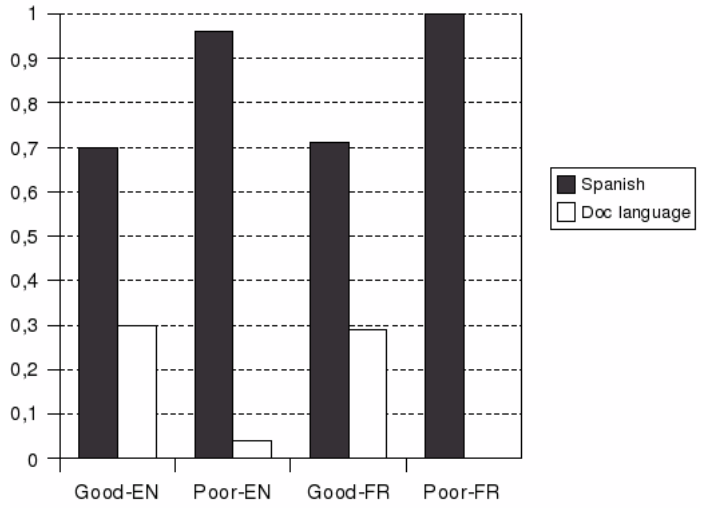

Figure 7: Input method at first stage of the search.

5.2 Query formulation and refinement

5.2.1 Initial search

In the first stage of the search, the user must choose the input method to formulate the question. Figure 7 shows the rate of queries launched in Spanish to be translated into document language before searching (translating&searching), versus queries launched directly in document language. 'Poor' groups scarcely used the latter (Poor-FR not at all).

'Good' groups formulated about 70% the question in Spanish. In these groups, all queries had individually similar rate, nevertheless, difference conduct was exhibited by users: each user usually used only one method for all questions. Many users of these groups reported that they used the first method (translating&searching) because were comfortable with it.

5.2.2 Refinement

When the answer was not found in the initial search, usually users reformulate the query. However, the refinement can be made in Spanish or in document language. Table 2 shows the average number of refinements per question after the initial search for both methods, and the average number of total searches per question (excluding initial search). No conclusions can be made showing this table: the

Group	Spanish	Doc language	# Refinements (av.)
Good-EN	0.23	0.98	1.20
Poor-EN	0.08	0.45	0.52
Good-FR	0.05	0.38	0.44
Poor-FR	0.27	0.23	0.49

Table 2: Query refinement.

Group	Refinement	Right	Not right
Good-EN	Not refined	40	26
	Refined	28	34
Poor-EN	Not refined	42	61
	Refined	6	19
Good-FR	Not refined	67	27
	Refined	19	15
Poor-FR	Not refined	62	35
	Refined	17	14

Table 3: Number of refined and not refined right assessed answers.

reason is that refinement depends highly on behaviour of each user. For example, user number 3 of Poor-EN group was the user who made the greater part of refinement in that group.

5.2.3 Quality of query translation

For a term driven document retrieval system, syntactic or grammatical quality of a translation has low importance. The translation of terms in correct context is important, and MT programs do it more o less good. When query was introduced or refined in Spanish, users only correct the translation when no answer was found. In the experiment, corrections was made in few cases: only 21 of a total of 251 (initial search plus refinement) Spanish-English translations, and only 21 of a total of 261 Spanish-French ones. Groups with good reading skills in document language made more corrections than 'Poor' groups. Users with good reading skills in French remarked that they were surprised with the quality of translated queries.

The number of right answers without refinement process was high. Table 3 shows the number of answers assessed right and not right.

5.3 Other aspects

5.3.1 Unsupported answer

The number of unsupported answer in English is very high. We think that it depends on two factors: the particular set of topics and the passage division, luckily worse for English collection than French one. If the context of the passage (i.e., the complete document) would be used, the right answers must be increased and inexact ones decreased.

5.3.2 'nil' answer

The average time for 'nil' answers was uniform for all tests: about 4 minutes, i.e., users abandon the search before time expired. It happens in the middle-end of the test, and denotes fatigue in the experiment (some users reported that test was a bit tired: many questions, some very large and complicate).

5.3.3 Topic number 9

The question "*¿Con el nombre de qué enfermedad se corresponde el acrónimo BSE?*" was judged different by English and French assessors. Nevertheless, for our tests with Spanish-English language pair difference

was very scarce. Only two answers (one for each Good-EN and Poor-EN groups) were necessary to modify: accuracy hardly varies.

6 Conclusions

The use of free on-line MT programs for the interactive CL-QA process has been explored in two important aspects: in searching process and in displaying information. In former, the number of right answer obtained with a MT version of the Spanish query is high. Difference exists with the question-document language pair: the Spanish-French was better than Spanish-English was. It is because of French is more closed language to Spanish than English is, and the quality of the translation between Spanish and French is better.

We expected that poor reading skills groups should obtain very much accuracy with the use of MT in displaying information, but difference is scarce when don't use it. In general, contrastive system obtains better appreciation, which is stronger for 'Poor' groups. However, these groups no obtain a significant improvement in accuracy with that system (even for the Poor-FR group the accuracy is worse). More experiment must be carried out to achieve other conclusions.

References

[1] C. G. Figuerola, Á. F. Zazo, J. L. Alonso Berrocal, and E. Rodríguez. REINA at the iCLEF 2004. In *Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK*, 2004.