Interactive Cross-Language Question Answering: Searching Passages versus Searching Documents

Fernando López-Ostenero, Julio Gonzalo, Víctor Peinado and Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos, UNED
{*flopez,julio,victor,felisa*}*@lsi.uned.es*

Abstract

iCLEF 2004 is the first comparative evaluation of interactive Cross-Language Question Answering systems. The UNED group has participated in this task comparing two strategies to help users in the answer finding task: the baseline system is just a standard document retrieval engine searching machine-translated versions of the documents; the contrastive system is identical, but searching passages which contain expressions of the appropriate answer type.

Although the users prefer the passage system because searching is faster and simpler, it leads to slightly worse results, because the document context (which is not available in the passage retrieval system) turns out to be useful to verify the correctness of candidate answers; this makes an interesting difference with automatic Q&A systems. In addition, our experiment sets a strong baseline of 69% strict accuracy, showing that Cross-Language Question Answering can be efficiently accomplished by users without using dedicated Q&A technology.

1 Introduction

In spite of its long name, "Interactive Cross-Language Question Answering" is not an exotic task, but rather a quite natural problem in the context of e.g. web searches: we want to know the answer to a question, and if the answer is out there in some web document, we want to find it as fast and easily as possible, and we do not want to miss the possibility of finding the answer just because it is written in a foreign language.

For our participation in the interactive CLEF track [3], which is for the first time devoted to study Cross-Language Question Answering (CL-QA) from a user inclusive perspective, we have designed an experiment aiming at:

- Establishing a reasonable baseline giving initial quantitative and qualitative data about the nature and difficulty of the task.
- Finding out whether passages are more adequate than full documents for interactive answer finding.
- Experimenting with interactive features specifically aimed at improving answer finding processes.

To achieve these goals, we have designed and compared two CL-QA assistants:

- A reference search system which uses a document retrieval engine (Inquery [2]) to retrieve machine-translated versions (into the user's native language) of the target language documents. Our hypothesis is that standard Document Retrieval and Machine Translation technologies, coupled together, can be efficient tools to help users in the answer location task.

- A contrastive search system which is identical to the reference system, except for two issues:
    1. It retrieves machine-translated passages rather than documents. The possibility of examining the context of a passage is intentionally excluded.
    2. At the beginning of the search, the user is asked to specify the type of answer (named entity, date, quantity), and only passages containing possible answers are retrieved and shown to the user.

In order to compare both systems, we have recruited eight Spanish native speakers with low English skills, which have searched the CLEF English collection to find answers for 16 questions extracted from the CLEF QA 2004 question set. Answers have been collected in Spanish, manually translated into English, and sent to CLEF QA assessors for evaluation (see [4] for details on the evaluation criteria).

Section 2 describes our experimental design, Section 3 discusses the results, and finally we draw some conclusions in Section 4.

2 Experiment design

Our experiment follows the iCLEF 2004 experiment design [1], which prescribes how to conduct searches with eight subjects, 16 fixed questions, fixed document collections for each available target language, and the two search systems being compared:

2.1 Test data

We have used the Spanish version of the question set, and the English text collection, which comprises news data from 1994 and 1995 taken from the Los Angeles Times and the Glasgow Herald. News were translated with Systran Professional 3.0 (as provided by the iCLEF organization).

2.2 User profiles

Our eight users were between 20 and 43 years old, had low or medium-low English skills, all were very familiarized with graphical interfaces and search engines, and in average they had little familiarity with Machine Translation systems.

2.3 Reference and Contrastive systems

Our **Reference system** is a straightforward document retrieval system (see Figure 1). Users type in queries in Spanish, and the system performs monolingual retrieval (using the Inquery API) against Systran translations of the original English news. The ranked list of results displays the document title, and the user can click to access the document contents. The interface has additional buttons to storage a document, to view storaged documents, and to end the search marking a document when an answer has been found.

The **Contrastive system** (Figure 2) begins by asking the user to select, from a radio button menu, which type of answer is appropriate for the question (a named entity, a date or a quantity). Then the search interface is similar, but

1. It retrieves machine-translated passages rather than documents. The possibility of examining the context of a passage is intentionally excluded, to test whether context is necessary or not to find and validate answers.

2. only passages containing the type of possible answers are retrieved and shown to the user.

The filter that discards unappropriate passages is straightforward and does not involve using any NLP tool:

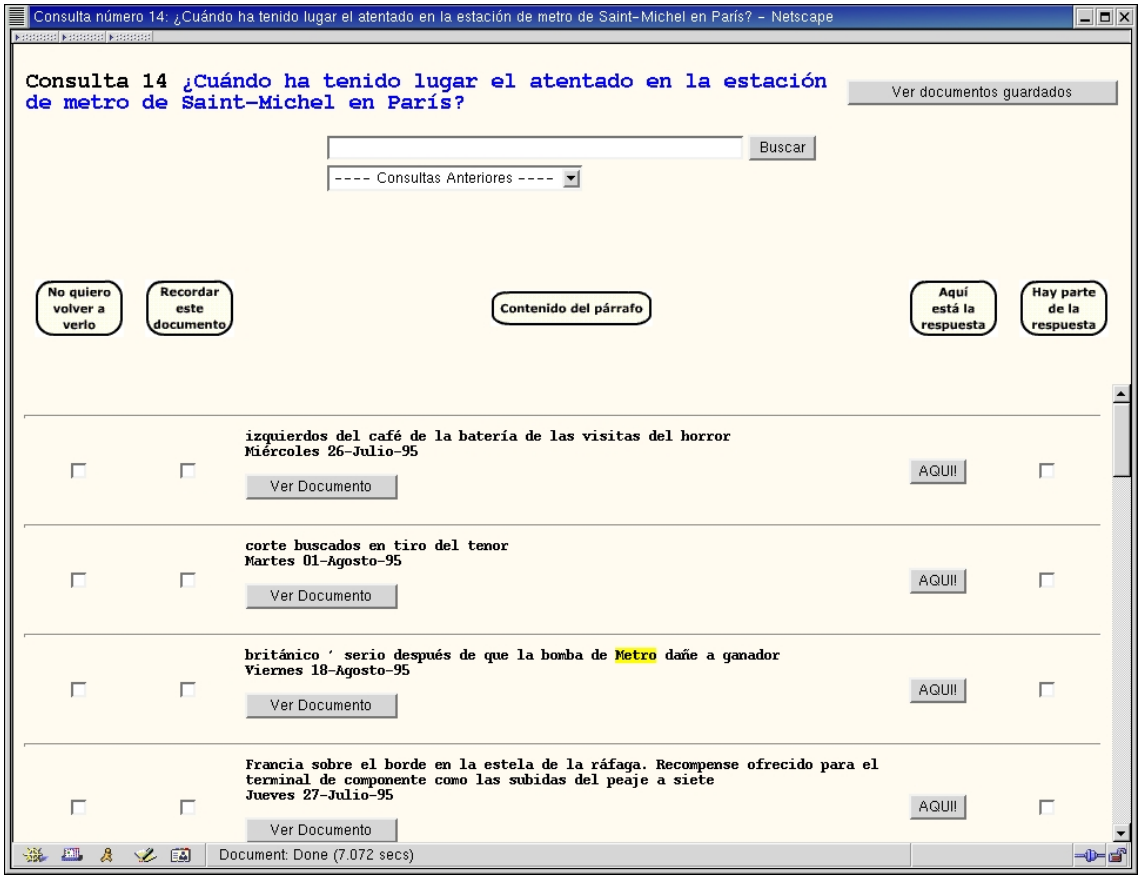

Figure 1: Documents system: Main interface

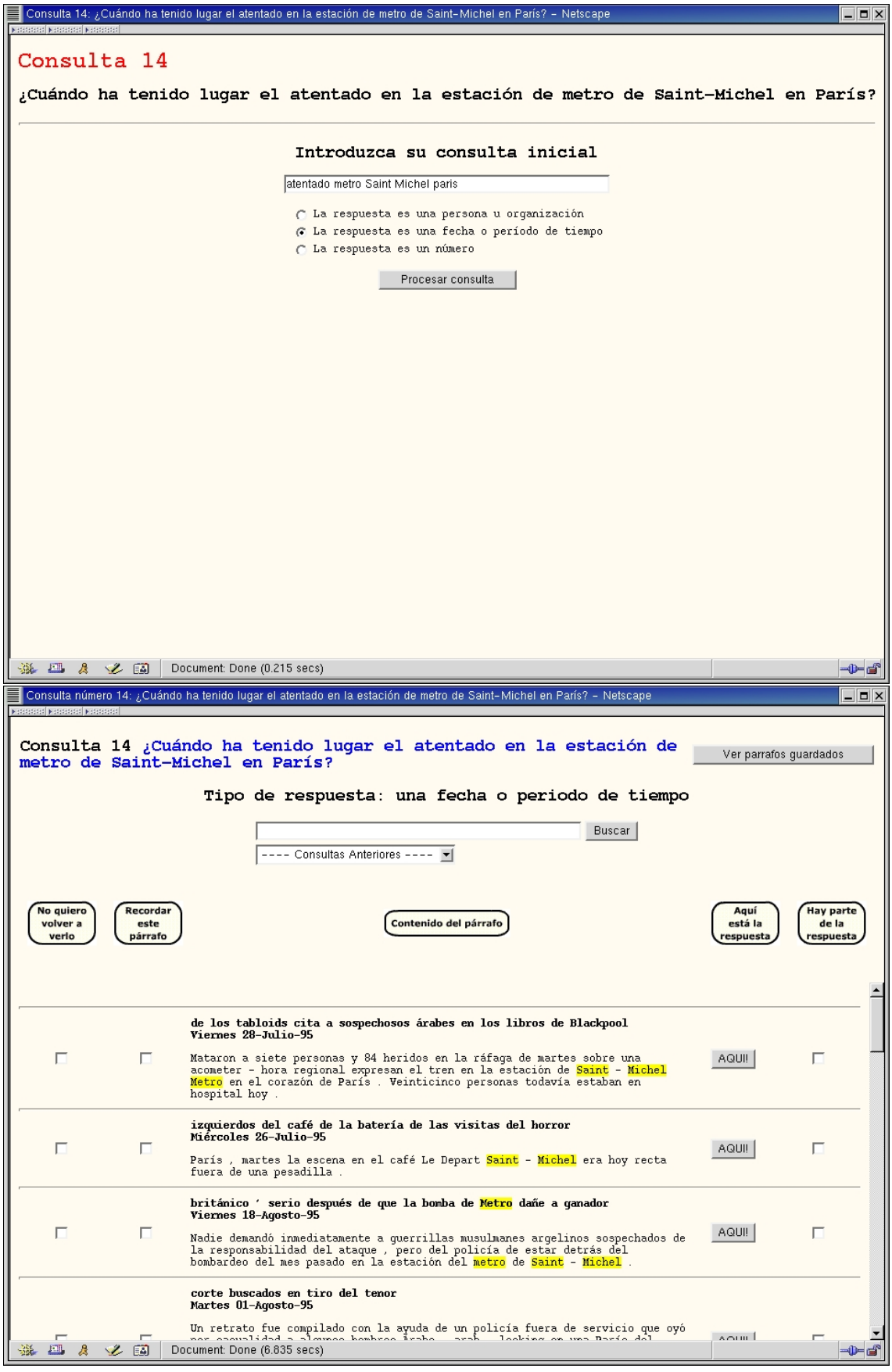

Figure 2: Passages system: Initial query and main interface

- A passage (sentences in our case) contains a named entity if there are expressions in uppercase where uppercase is not prescribed by punctuation rules. Locations are looked up in a gazeteer and filtered out, because "location" questions are excluded from the iCLEF question set.
- A passage contains a temporal reference if there is a match with a list of words denoting dates or a number between 1900 and 1999 (this temporal restriction is ad-hoc for CLEF data).
- Similarly, a passage contains a quantity if there is a number or a word from a given list.

Note that the aim of the filter is deciding whether there are named entities, quantities or dates, not finding them: that makes the task much easier. Note also that recall is much more important than precision, because we do not want to miss any potential answer. That makes our simple filter effective for our purposes, and its potential mistakes relatively harmless.

Overall, the filter identifies named entities in 75% of the sentences, which is too permisive to be useful. A real Named Entity Recognizer, able to distinguish between people, organizations, locations, etc., would be a useful substitute of our naive filter. In the other two categories, however, the filter is useful: only 21% of the sentences contain dates, and 43% contain quantities.

2.4 Search sessions

Every subject searches all 16 questions, eight with each system, according to the latin-square matrix design prescribed by the iCLEF guidelines (see Table 1). They filled in a pre-search questionnaire, two post-system questionnaires, and a final post-search questionnaire. The maximum search time per question was five minutes. Once time expired, the system stops the search, and the user has a last chance of writing an answer.

user	**search order** (system: $A\|B$, question: $1 \ldots 16$)
1	A1 , A4 , A3 , A2 , A9 , A12, A11, A10, B13, B16, B15, B14, B5 , B8 , B7 , B6
2	B2 , B3 , B4 , B1 , B10, B11, B12, B9 , A14, A15, A16, A13, A6 , A7 , A8 , A5
3	B1 , B4 , B3 , B2 , B9 , B12, B11, B10, A13, A16, A15, A14, A5 , A8 , A7 , A6
4	A2 , A3 , A4 , A1 , A10, A11, A12, A9 , B14, B15, B16, B13, B6 , B7 , B8 , B5
5	A15, A14, A9 , A12, A7 , A6 , A1 , A4 , B3 , B2 , B5 , B8 , B11, B10, B13, B16
6	B16, B13, B10, B11, B8 , B5 , B2 , B3 , A4 , A1 , A6 , A7 , A12, A9 , A14, A15
7	B15, B14, B9 , B12, B7 , B6 , B1 , B4 , A3 , A2 , A5 , A8 , A11, A10, A13, A16
8	A16, A13, A10, A11, A8 , A5 , A2 , A3 , B4 , B1 , B6 , B7 , B12, B9 , B14, B15

Table 1: iCLEF 2004 Latin-Square Experiment Design

3 Results and discussion

System	Accuracy		Time	Confidence		# Refinements
	strict	lenient	(av.)	High	Low	(av.)
Documents	.69	.73	209.05	44	20	1.6
Passages	.66	.72	195.20	41	23	1.7

Table 2: Comparison of results for both systems

3.1 Comparison between systems

The main differences (in search results and search behaviour) between systems can be seen in Table 2. The average (strict) accuracy is 5% higher for the baseline system, and the search behaviour (average searching time, confidence in the answers, average number of refinements) is very similar for both systems. The absolute performance (between .66 and .69 strict accuracy) is

remarkably high: there is room for improvement, but it is fair to say that users can find answers efficiently without QA-specific machinery. This accuracy is obtained in an average time of only $3,5$ minutes, and with only 1.6 average refinements per question.

Why our contrastive system, which has some QA-specific features, performs worse than the baseline system? Our observational studies, together with the questionnaires filled by our users, give some hints:

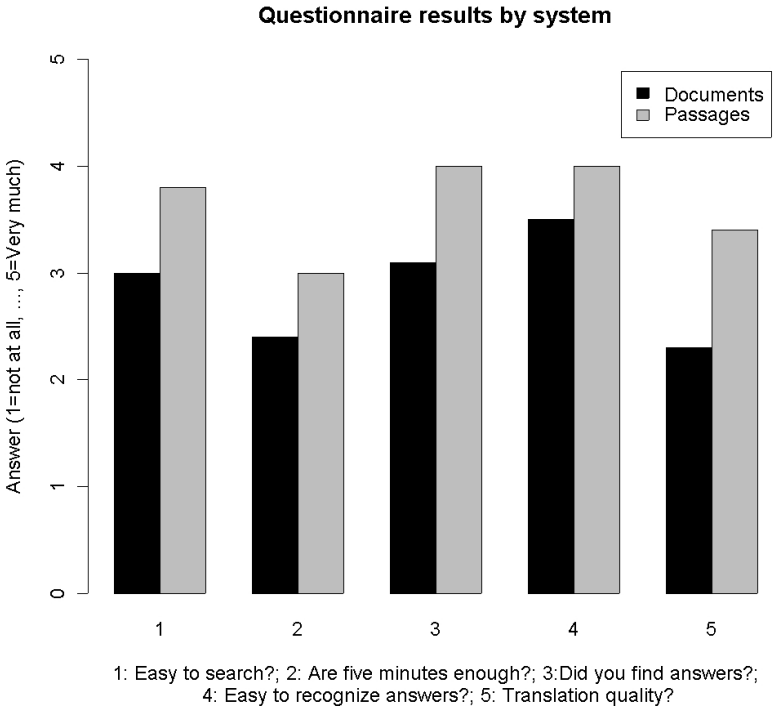

Figure 3: Post-system questionnaires

- The results of post-system questionnaires (where users evaluated each system separately) can be seen in Figure 3. For all individual questions, the passage-retrieval system had better results: according to users, it was easier to search with, faster, it was easier to recognize answers, and even the translation quality (which is the same) was perceived as better. We can conclude that users felt more comfortable when searching with the contrastive system. Then why the performance is worse?

- The results of the post-search questionnaire (where users were explicitly instructed to compare systems) can be seen in Figure 4. In this explicit comparison, again the passage-retrieval system is perceived as "easier to learn" and "easier to use". But when asked for the better system overall, both systems receive half of the votes. Why?

- The written comments made by our subjects, together with our observational study, give a clear answer: most subjects wrote that the passage retrieval system was easier and faster to use, but only the document retrieval system permitted looking at the full content of documents containing potential answers; and the context was perceived as a key factor to ensure that a potential answer was correct. In addition, the document context was also used to refine the query and/or search for similar documents that might verify a potential answer. This is a factor related not only to a document content, but also to the translation quality, which often creates doubts about the correctness of an apparent answer.

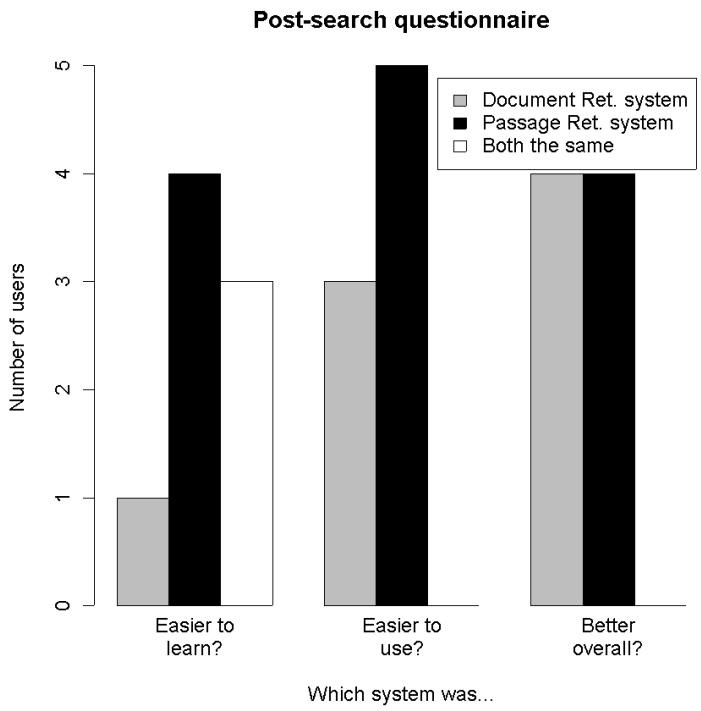

Figure 4: Post-search questionnaire comparing systems

From the comments made by our subjects, and from their search behaviour, it seemed clear that the preferred search facility would do passage retrieval, but with the possibility of accessing the context of a passage when desired.

3.2 Failure analysis

Out of 148 answers, there were 33 judged as "W" (wrong), 7 as "inexact" and 1 as "unsupported". 21 wrong answers were simply time outs: the user was not able to find an answer in five minutes. In the remaining $12 + 7 + 1 = 20$ cases, users gave an answer that was not correct. The sources of error are:

Misleading translations In most cases, an incorrect or misleading translation is the responsible for an incorrect answer. In some cases, users were doubtful about an answer, and looked for additional evidence supporting the answer. Once time has expired, they preferred to give an answer with a low level of confidence, than no answer at all.

Human errors In a few cases, the user just made a mistake when writing the answer. For instance, a user stated that the Channel Tunnel costed "15,000 millions", without specifying the currency. In other occasion, a cut-and-paste error repeated the answer given for a former question.

Responsiveness criteria Occasionally, the user and the assessor had different opinions about the responsiveness or focus of an answer. For instance, when asking for the number of missing people caused by the typhoon Angela, a user said "more than 500 killed and 280 still missing", which was judged as inexact.

It is worth noticing that, while automatic Q&A systems may avoid translating documents (by translating only selected query terms), in an interactive system it is unavoidable to translate

documents if the user does not have target-language skills. Thus, accurate targeted translation is a specific requirement of interactive CL-QA systems.

It is also worth noticing that, in some occasions, users were able to jump over significant translation problems. For instance, *When did Latvia gain independence?* was pretty hard to answer, because Systran did not have "Latvia" in its bilingual vocabulary; thus, it remained untranslated in all documents. Users were looking for "Letonia" (Spanish translation of Latvia), but nothing was found. Some documents, however, spoke about "Latvia, Estonia y Lituania", and users were able to "disambiguate" Latvia from the context.

It is also worth mentioning that users were able to make more inferences than current Q&A systems. An interesting example is *When did Lenin die?* Some users found a document talking about the beginning of the celebrations of the 70^{th} anniversary of Lenin's death. Substracting from the date of the document (Sunday 22 January, 1994), users correctly deduced 1924. A couple of users, however, answered "20 January, 1924" (because the document asserts that "celebrations started last Friday") which was incorrect, because the celebrations started on January 20^{th} but the real anniversary was on January 21^{st}.

3.3 User effects

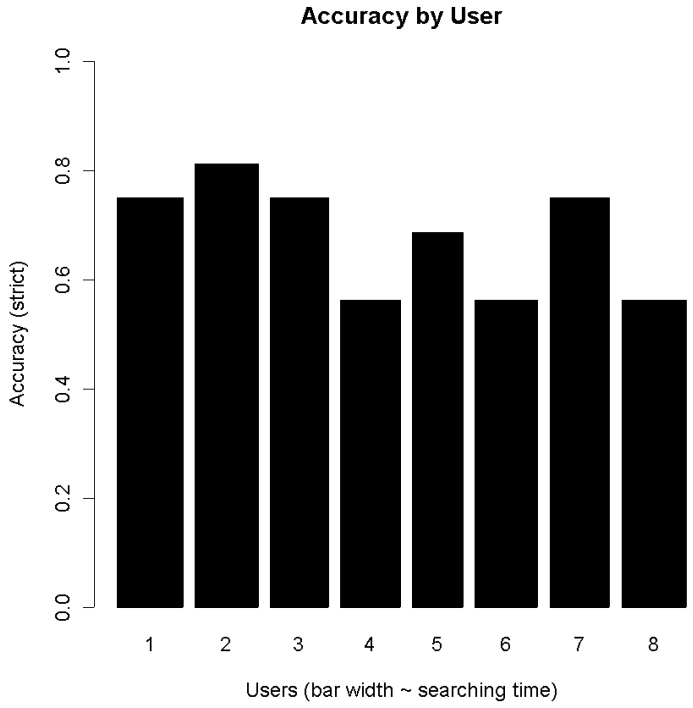

Figure 5: Results by user

The accuracy by user, and its correlation with average search time, can be seen in Figure 5. Both accuracy and average search time are rather homogeneous across users; more than in our previous experiences in interactive CL Document Retrieval. Our impression is that users find the Q&A task simpler to understand, easier and more amusing than document retrieval; thus, fatigue effects are less relevant. In some cases, a priori knowledge of the question domain permitted a better selection of query terms, but the effect on average accuracy is small.

3.4 Topic effects

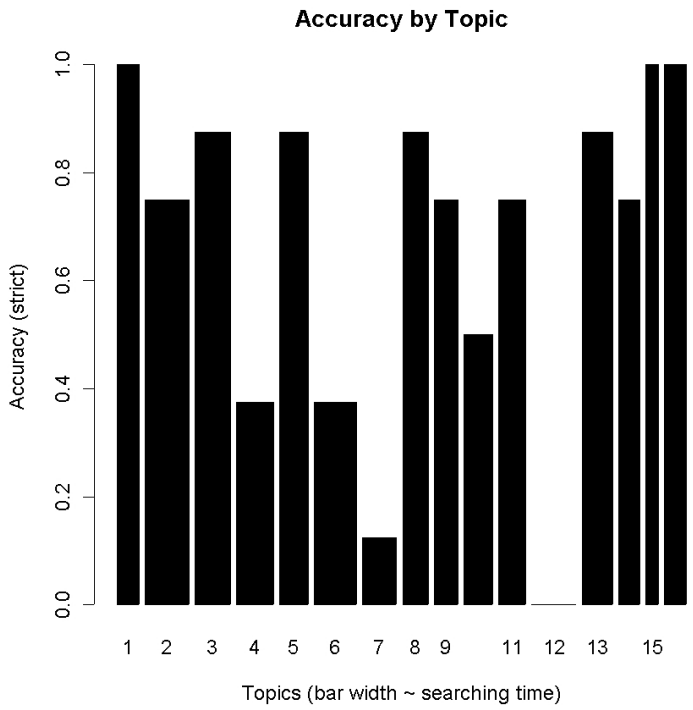

Figure 6: Results by topic

The accuracy by topic, and its correlation with average search time, can be seen in Figure 6. Obviously, the effect of topic difficulty is the predominant factor on accuracy. The most difficult topics are:

- *What is Charles Millon's political party?* (accuracy 0). No reference to Charles Millon was found by any user, probably due to mistranslation by Systran.
- *How many people were declared missing in the Philippines after the Thyphoon "Angela"?* (accuracy 1/8). The source of difficulty here is not in translation quality. Some users answered with preliminary, vague information; others mixed dead and dissapeared people.
- *When did Latvia gain independence?* (accuracy 3/8). As commented above, the source of difficulty is that Systran left "Latvia", which is the key term in this question, untranslated.
- *Who committed the terrorist attack in the Tokyo underground?* (accuracy 3/8). Apparently the source of difficulty was cross-linguality: it was hard to find good query terms, specially because "Tokyo" was misspelled in the Spanish question (it used the English spelling, Tokyo, instead of the Spanish spelling, Tokio) and Systran mixed both spellings.

4 Conclusions

Our first experiment in interactive Cross-Language Question Answering has produced some interesting results:

- First, we have set a strong baseline (69% strict accuracy) for the task, using standard Document Retrieval and Machine Translation technologies. Can automatic CL-QA systems

be adapted to interactive settings to achieve significantly higher user performance? This is an interesting research question for upcoming iCLEF editions.

- The main source of difficulty is the cross-language nature of the search, rather than the idiosincrasy of the QA task. Task-specific term suggestion and machine translation techniques might be useful for interactive CL-QA.
- Users prefer passage retrieval to document retrieval for the CL-QA task, partly because full Machine Translated documents are noisy and discouraging. But once a potential answer is found, the context is sometimes helpful to validate it.
- Interactivity can effectively be used to add Q&A specific restrictions to focus a passage search. In our contrastive system, users were asked to specify which type of answer was required for the question at hand. When, for instance, the answer had to be a date, only 21% of the sentences in the collection had to be searched.

Acknowledgements

This research has been partially supported by a grant from the Spanish Government, project R2D2 (TIC2003-07158-C04-01), and a grant by the UNED (Universidad Nacional de Educación a Distancia).

References

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

[2] J. Callan, B. Croft, and S. Harding. The INQUERY retrieval system. In *Proceedings of the 3rd Int. Conference on Database and Expert Systems applications*, 1992.

[3] J. Gonzalo and D. Oard. iCLEF 2004 Track Overview: Pilot Experiments in Interactive Cross-Language Question Answering. In *This volume*, 2004.

[4] B. Magnini, A. Vallin, C. Ayache, M. Rijke, G. Erbach, A. Peñas, D. Santos, K. Simov, and R. Sutcliffe. Overview of the CLEF 2004 Multilingual Question Answering Track. In *This volume*, 2004.