UNED at iCLEF 2005: Automatic highlighting of potential answers

Víctor Peinado, Fernando López-Ostenero, Julio Gonzalo and Felisa Verdejo
NLP Group, ETSI Informática, UNED
c/ Juan del Rosal, 16, E-28040 Madrid, Spain
{victor, flopez, julio, felisa}@lsi.uned.es

Abstract

In this paper, we describe UNED's participation in the iCLEF 2005 track. We have compared two strategies for finding an answer using an interactive question answering system: i) a search system over full documents and ii) a search system over passages (document's paragraphs). We have added an interesting feature to both system in order to facilitate reading: the possibility to enable/disable the highlighting of named entities such as proper names, temporal references and numbers likely to contain the right answer.

Our Document Searcher obtained better overall accuracy (.53 vs. .45) but our subjects found browsing passages simpler and faster. However, most of them presented a similar search behavior (regarding time consumption, confidence in their answers and query refinements) using both systems. All our users considered helpful the highlighting of named entities and they all made extensive use of this possibility as a quick way of discriminating between relevant and non relevant documents and finding a valid answer.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.4 [**Information Systems Applications**]: H.4.m Miscellaneous

General Terms

highlighting, interactive QA

Keywords

interactive question answering, named entities recognition, cross-language information retrieval, search behavior

1 Introduction

The main goal of the Interactive Cross-Language Question Answering task (iCLEF) consists of finding an answer for 16 general questions (e. g. *Who is the president of Burundi?*) and selecting a certain document that supports the answer, before the time limit of five minutes expires.

Our participation in iCLEF 2004 [4] focused on comparing two strategies for finding an answer using an interactive question answering (QA) system: **i)** a documents retrieval search engine and;

ii) a passages retrieval search engine. We wanted to study what approach was more helpful: browsing documents or passages?

Our subjects preferred the passages system because browsing paragraphs was simpler and faster, but they also missed the possibility of accessing the full context of the passage since sometimes it was difficult to understand the context of the paragraph. But, in spite of the preferences, average strict accuracy turned out to be slightly higher in the documents system (69%).

This year we intended to study the impact of automatic highlighting of named entities in both systems. First of all, in the Passages system, we allowed our subjects to visualize the full contents of the documents. Then, we made use of our simple recognizer, which was able to locate proper nouns, temporal references and numbers, and we added the possibility of enable and disable the emphasis of these named entities. Is it helpful to highlight the named entities in order for the subjects to find a possible answer? How much does the highlighting help the user while browsing documents and while browsing passages?

The remaining sections of this paper are divided as follows. In Section 2, we describe the design of the experiments, our testbed and how search sessions are organized. In Section 3, we present our two cross-language search systems. Then, in Section 4, we discuss the official results, analyzing the causes of failure (4.2), the users' and topics' effects (4.3 and 4.4) and the cases in which subjects found the answer in the Passages system thanks to the possibility of access the full document (4.5). Lastly, in Section 5, we present some conclusions.

2 Experiment design

2.1 Testbed

Following the iCLEF 2005 guidelines, [1] we have carried out the comparison of two different cross-language search systems. Eight subjects have searched for the answer of 16 fixed questions in Spanish over a collection of documents written originally in English. The subjects performed eight queries with each system, according to the design of a latin-square proposed by the organization of the task [3].

The collection of documents consisted of news from 1994 and 1995 taken from *Los Angeles Times* and *Glasgow Herald* newspapers, respectively. In our experiments, we did not use the original documents but a Spanish version translated with *Systran Professional 3.0.*

From this translated version of the collection, we made use of the Inquery's API [1] in order to build two different indexes, one for each search system:

1. One index whose documents correspond with news articles.

2. Another one in which each document corresponds with a single passage (a paragraph of a news article).

We recruited eight users who were between 19 and 30 years old and had different levels of education, from high school to master degrees. Their mother tongue was Spanish and they all claimed to have between low and medium-high skills in written English comprehension. They were highly familiarized with graphical interfaces and web-based search engines. They also declared to have been using WWW search engines for at least 2-7 years (avg=4.6). On the contrary, none of them had any familiarity using Machine Translation (MT) systems.

2.2 Search sessions

We asked the subjects to find a valid answer and select a document supporting it before the time limit. The maximum search time per question was set in five minutes. Once time expired, the system stopped the search and allowed to visualize the subject the set of stored documents, giving her/him a last chance to write an answer.

[1] For further details, please see `http://nlp.uned.es/iCLEF`.

They also had to fill in a pre-search questionnaire about their previous experience with search engines, two post-system questionnaires analyzing their performance and the specific features of each approach, and a final post-search questionnaire about their overall experience.

3 Description of the reference and contrastive systems

3.1 Reference system

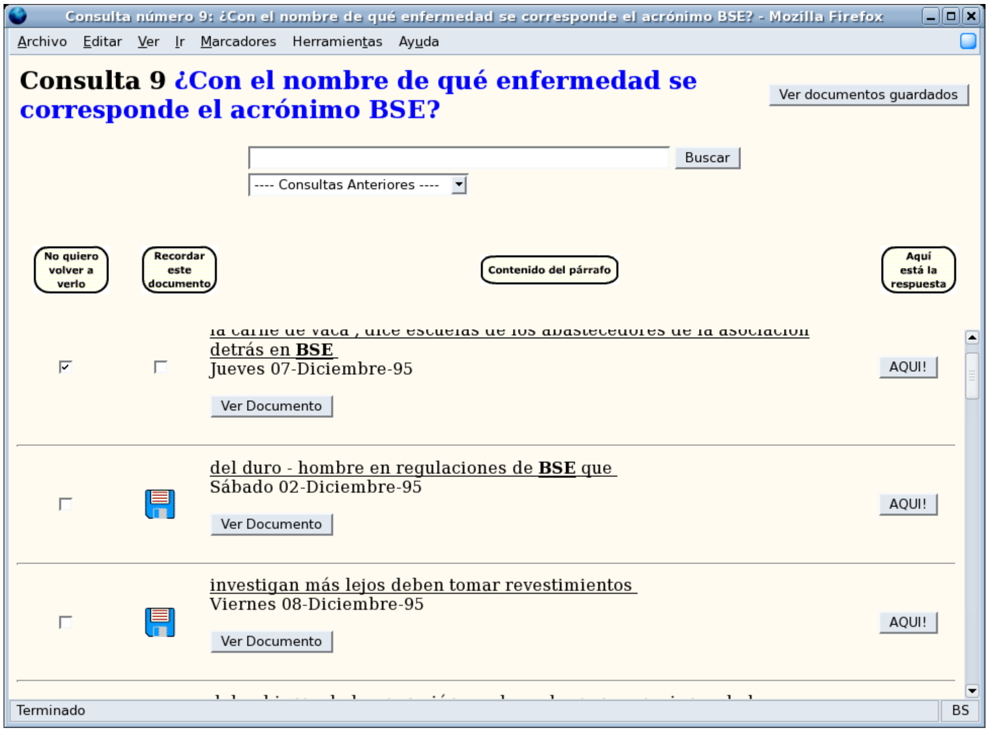
Our reference system, henceforth the **Documents Searcher**, is a simple traditional search engine in which each retrieved document corresponds with a complete news article. Indeed, it has few differences compared to the reference system used last year [4].

We may outline the normal sequence of a subject's actions as follows:

1. The subject **types the query terms** in Spanish and launches the query.

2. The system makes use of the Inquery's API to retrieve a **ranking of relevant documents**.

3. The **main interface** displays only the titles and dates of each document (see Figure 1). This interface has additional buttons to discard non-relevant documents, to store a certain document considered interesting, to list already stored documents, and to conclude the search selecting a certain document when an answer has been found.


Consulta número 9: ¿Con el nombre de qué enfermedad se corresponde el acrónimo BSE? - Mozilla Firefox
Archivo Editar Ver Ir Marcadores Herramientas Ayuda
Ver documentos guardados
Buscar
---- Consultas Anteriores ----
No quiero volver a verlo
Recordar este documento
Contenido del párrafo
Aquí está la respuesta
AQUI!
Ver Documento
Terminado
BS

Consulta 9 ¿Con el nombre de qué enfermedad se corresponde el acrónimo BSE?

☑ ☐ la carne de vaca , dice escuelas de los abastecedores de la asociacion detrás en **BSE**
Jueves 07-Diciembre-95

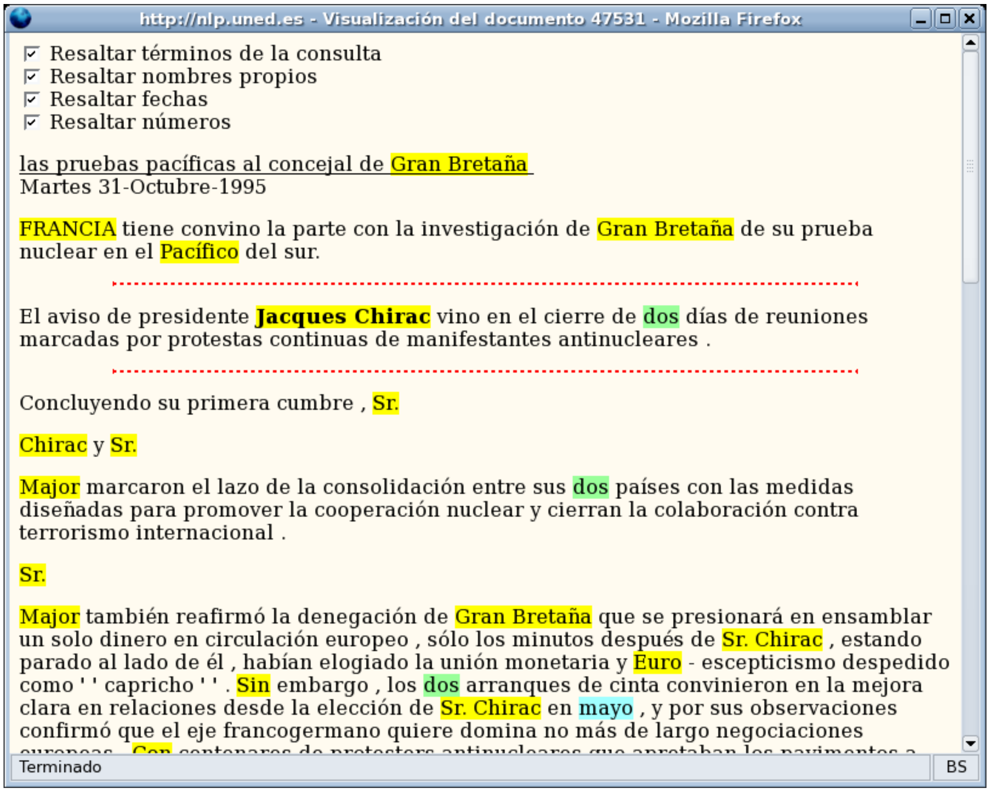
del duro - hombre en regulaciones de **BSE** que
Sábado 02-Diciembre-95

☐ investigan más lejos deben tomar revestimientos
Viernes 08-Diciembre-95

Figure 1: Documents Searcher's main interface.

4. From this main interface, it is possible to **visualize the whole document**.

   We have added a feature that did not exist in last year's systems in order to improve the reading: query terms' occurrences appear within the text in boldface. In addition, it is possible to handle some checkboxes in order to enable/disable the highlighting of named entities, such as proper nouns, temporal references, dates and numbers. See Figure 2 for a detailed screenshot showing the highlighting.


http://nlp.uned.es - Visualización del documento 47531 - Mozilla Firefox
Terminado
BS

☑ Resaltar términos de la consulta
☑ Resaltar nombres propios
☑ Resaltar fechas
☑ Resaltar números

las pruebas pacíficas al concejal de Gran Bretaña
Martes 31-Octubre-1995

FRANCIA tiene convino la parte con la investigación de Gran Bretaña de su prueba nuclear en el Pacífico del sur.

El aviso de presidente **Jacques Chirac** vino en el cierre de dos días de reuniones marcadas por protestas continuas de manifestantes antinucleares .

Concluyendo su primera cumbre , Sr.

Chirac y Sr.

Major marcaron el lazo de la consolidación entre sus dos países con las medidas diseñadas para promover la cooperación nuclear y cierran la colaboración contra terrorismo internacional .

Sr.

Major también reafirmó la denegación de Gran Bretaña que se presionará en ensamblar un solo dinero en circulación europeo , sólo los minutos después de Sr. Chirac , estando parado al lado de él , habían elogiado la unión monetaria y Euro - escepticismo despedido como ' ' capricho ' ' . Sin embargo , los dos arranques de cinta convinieron en la mejora clara en relaciones desde la elección de Sr. Chirac en mayo , y por sus observaciones confirmó que el eje francogermano quiere domina no más de largo negociaciones europeas . Con centenares de protestors antinucleares que apretaban los pavimentos a

Figure 2: Highlighted named entities: query terms in boldface, proper nouns in yellow, temporal references in blue and numbers in green.

5. Lastly, the subject must **type the answer and assign it a confidence value**: high or low.

3.2 Contrastive system

We propose as contrastive system a **Passages Searcher**, which performs the queries over a collections of news paragraphs.

In this case, the sequence of actions is the following:

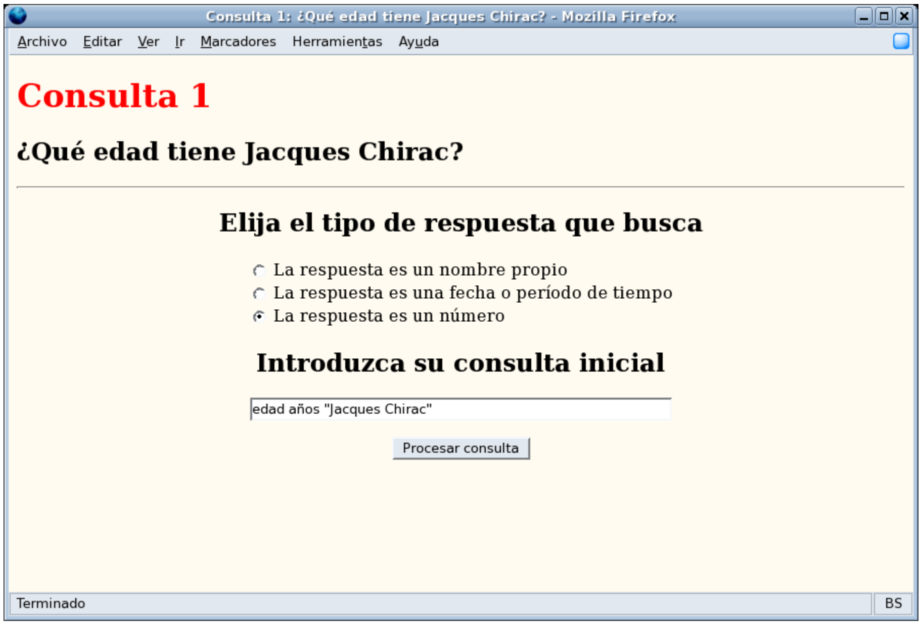
1. First of all, the subject is asked to **choose the type of answer** she/he is searching for: a proper noun, a date or a number (see Figure 3).

   Notice that: **i)** this distinction agrees with the three different types of named entities identifiable by our recognizer[2] and; **ii)** this initial choice determines which pieces of information will be automatically highlighted.

   The underlying idea is that, in order to facilitate reading and locating a possible answer, the system will highlight named entities of the same type of the one chosen before submitting the query. For instance, if a subject if looking for a date, it can be useful to automatically emphasize all kind of temporal references.

2. The subject **types the query terms** in Spanish and launches the query.

[2] We have used a straightforward recognizer which is able to identify proper nouns, temporal references and numbers. See also [5].



Consulta 1: ¿Qué edad tiene Jacques Chirac? - Mozilla Firefox
Archivo Editar Ver Ir Marcadores Herramientas Ayuda
Procesar consulta
Terminado
BS

Consulta 1

¿Qué edad tiene Jacques Chirac?

Elija el tipo de respuesta que busca

○ La respuesta es un nombre propio
○ La respuesta es una fecha o período de tiempo
● La respuesta es un número

Introduzca su consulta inicial

edad años "Jacques Chirac"

Figure 3: Submitting the query in the Passages Searcher.

3. The system retrieves and shows a **ranking of relevant passages**. Those passages containing the selected type of answer are promoted by the search engine, and the system automatically highlights query terms and named entities, depending on the initial subject's election.

4. The **main interface**, as shown in Figure 4, provides also titles and dates of each news article, and has the same buttons that the Documents Searcher to discard and store documents.

   Unlike last year's experiments, now it is possible to access the complete document the passage makes part of. If this situation takes place, the whole document will clearly show the passage with two dashed lines.

   In our participation in iCLEF 2004[4], we intentionally excluded the possibility of examining the context of a given passage by providing the complete document. All our subjects expressed their complaints because this lack hindered them from understanding the general sense of some short paragraphs. In addition, other works had already analyzed the benefits of allowing the subjects to get the full contents of the documents [2] and we decided to add this feature.

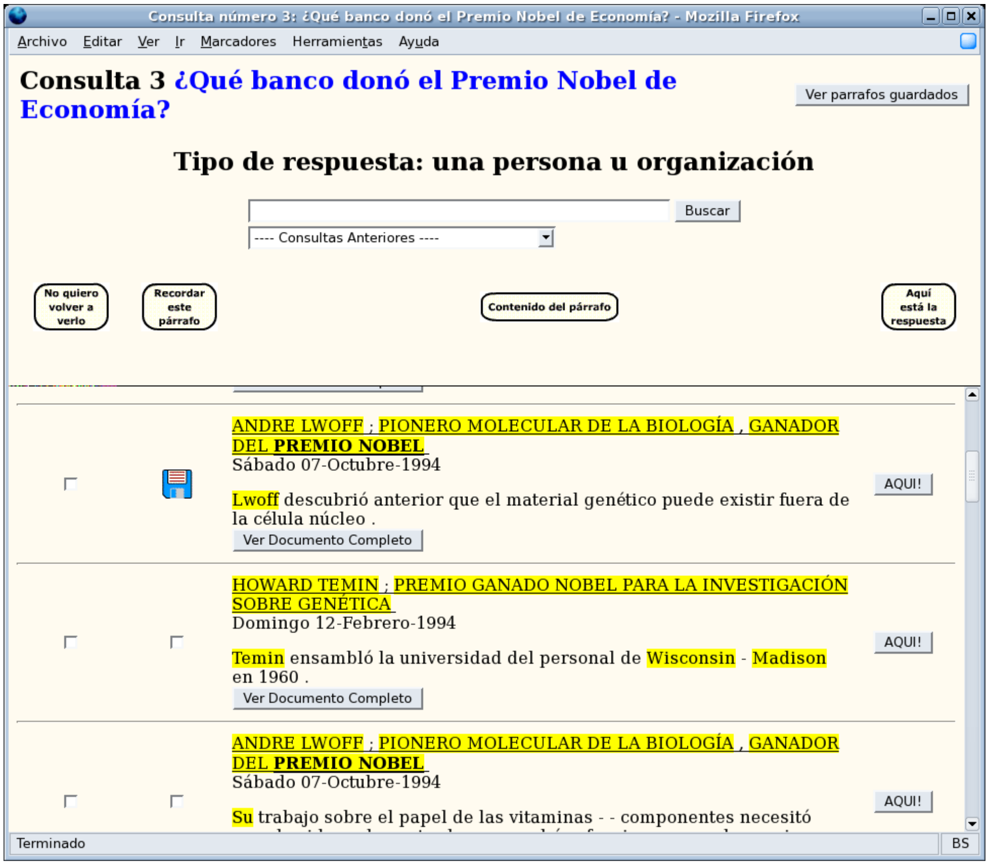
5. As in the Documents Searcher, when **visualizing the full document**, it is possible to enable/disable the highlighting of query terms, proper nouns , temporal references and numbers (Figure 2).

6. Lastly, the subject must **type the answer and assign it a confidence value**: high or low.

4 Results and discussions

4.1 Comparison between systems

From the general results shown in Table 1, we can remark the following:



Consulta 3 ¿Qué banco donó el Premio Nobel de Economía?

Tipo de respuesta: una persona u organización

□ ANDRE LWOFF ; PIONERO MOLECULAR DE LA BIOLOGÍA , GANADOR DEL PREMIO NOBEL
Sábado 07-Octubre-1994

Lwoff descubrió anterior que el material genético puede existir fuera de la célula núcleo .
Ver Documento Completo
AQUI!

□ □ HOWARD TEMIN ; PREMIO GANADO NOBEL PARA LA INVESTIGACIÓN SOBRE GENÉTICA
Domingo 12-Febrero-1994

Temin ensambló la universidad del personal de Wisconsin - Madison en 1960 .
Ver Documento Completo
AQUI!

□ □ ANDRE LWOFF ; PIONERO MOLECULAR DE LA BIOLOGÍA , GANADOR DEL PREMIO NOBEL
Sábado 07-Octubre-1994

Su trabajo sobre el papel de las vitaminas - - componentes necesitó
AQUI!

Figure 4: Passages Searcher's main interface.

System	Accuracy		Time	Confidence		Refinements
	strict	lenient	(avg)	High	Low	(avg)
Documents	.53	.53	222.25	36	28	2.42
Passages	.45	.45	220.77	36	28	2.28

Table 1: Comparison of results for both systems

1. The Documents Searcher obtained again better accuracy than the Passages Searcher: .53 and .45, respectively.

2. Both systems got the same values of strict and lenient accuracy. None of our subject's answers was judged as inexact by the assessors.

3. Regarding the average time consumption, confidence values and the average number of refinements, our subjects present a quite similar behavior with both systems.

The 2004 and 2005 results are not directly comparable because the topics, the systems' features, the participating subjects and the conditions of the experiments were not obviously the same. Nevertheless, the difference between the two strategies has increased: now the Passages Searcher has been 15% worse than the Documents Searcher.

4.2 Failure analysis

Most of the failure causes was related to mistranslations. As we will discuss below in Section 4.4, in some occasions, the MT system did not translated correctly, for instance, translating some terms when it shouldn't and vice versa.

There were also remarkable human errors. Specifically, some users got confused in those topics in which different potential answers (some of them looking contradictory) appeared in the collection (e.g. topics asking for a number of casualties in a incident).

Regarding responsiveness criteria, the results have been strongly language-biased because the same answer was judged in a different way by English and French assessors (see Section 4.4).

4.3 User effects

User	**Accuracy**		**Confidence**				**Refinements**		**Time**	
	Docs	**Pass**	**Docs**		**Pass**		**Docs**	**Pass**	**Docs**	**Pass**
			High	Low	High	Low				
1	.62	.50	4	4	5	3	3.88	2.75	280.36	238.63
2	.25	.50	3	5	5	3	3.5	2.36	275.13	240.75
3	.62	.38	6	2	6	2	1.62	1.36	199.63	197.25
4	.50	.38	4	4	3	5	1.36	1.36	187.88	240.88
5	.75	.62	4	4	5	3	1.75	1.36	251.25	179.75
6	.25	.75	5	3	6	2	1.75	2.25	201.75	179
7	.62	.12	5	3	3	5	2.36	3.38	148.88	245.38
8	.62	.38	5	3	3	5	3.12	3.38	233.13	244.5

Table 2: **Accuracy, confidence, refinements and time (in seconds) per user**

The data about accuracy, confidence, number of refinements and time consumption per user are shown in Table 2. Seven out of the eight subjects stated in the questionnaires that they preferred the Passages Searcher. However, six out of eight found more right answers with the Documents Searcher. Some users had some difficulties when using one of the systems. User 7, particularly, obtained poor results with the Passages Searcher, in spite of the fact that he spent, on average, 245.38 seconds for each topic. On the contrary, users 2 and 6 performed much worse with the Documents searcher.

Notice that confidence values are generally coherent with the accuracy. Except for users 3 and 6, there are no big differences between the number of answers with a high confidence and the accuracy. For instance, user 6 assigned a high confidence to five of the topics performed with the Documents Searcher but obtained an accuracy of .25, representing only two answers assessed as right.

Also, there seems to be a certain correlation between number of query refinements and the experience using our systems, because the three subjects who had already collaborated in 2004 (3, 5, 6) made, on average, fewer refinements than the others.

4.4 Topic effects

Table 3 shows values about accuracy, confidence, refinements and time consumption per topic. The data clearly pinpoint the difficulties of finding the correct answer for some topics. Those topics in which our subjects obtained poor accuracy, made more refinements and spent longer are:

- 12: *When do we estimate that the Big Bang happened?* In the astronomic domain, the English term "Big Bang" is used as is in Spanish but in our collection it had been translated as *"Gran Estallido"*. This misled most of our subjects and only one of them was able to find a valid answer.

Topic	**Accuracy**		**Confidence**				**Refinements**		**Time**	
	Docs	**Pass**	**Docs**		**Pass**		**Docs**	**Pass**	**Docs**	**Pass**
			High	Low	High	Low				
1	.75	.50	2	2	3	1	3	2.5	270.75	233.75
2	.75	.50	3	1	2	2	2	2	227.25	249.75
3	.25	.50	1	3	3	1	4	2	288	242.5
4	.50	0.00	2	2	1	3	1.5	2.75	203	243.75
5	.25	.50	1	3	2	2	3.75	3	278.25	300
6	.75	.75	3	1	3	1	.75	1.25	268.75	227.5
7	.75	.25	4	0	3	1	.25	1.75	179.75	229.5
8	1.00	1.00	4	0	3	1	1.25	0	126.25	87.25
9	.50	.25	4	0	4	0	.25	.25	73.25	114.25
10	1.00	1.00	3	1	4	0	1.75	1	167	132.25
11	.75	1.00	2	2	4	0	3.25	.25	211.75	160.25
12	.25	0.00	1	3	0	4	4.5	6.5	300	300
13	0.00	0.00	1	3	0	4	5.5	4.5	300	294.75
14	0.00	0.00	0	4	0	4	2.5	3.5	300	300
15	.25	0.00	2	2	0	4	3.5	4	253.25	300
16	.75	1.00	3	1	4	0	1	1.25	108.75	116.75

Table 3: **Accuracy, confidence, refinements and time (in seconds) per topic**

- 13: *Who won the Miss Universe 1994 beauty contest?* As in the previous topic, here there was a translation problem. "Miss Universe" was only partially translated and abbreviated as *"Srta. Universe"* instead of the correct translation that should have been *"Miss Universo"*. Besides, it became complicated even to find a document related to this beauty contest.

- 14: *How many countries have ratified the United Nations convention adopted in 1989?* What made difficult to find a valid answer for this topic was perhaps the huge number of documents related to countries ratifying UN conventions. None of our subjects was able to find a right document with the correct answer.

- 15: *How many states are members of the Council of Europe?* Most of our subject misunderstood the Council of Europe with the European Union.

Topic 9 (*What disease name does the acronym BSE stand for?*) was thought to be an easy topic and its low accuracy deserves a more detailed explanation. While English assessors considered with good sense that answers different from "Bovine Spongiform Encephalopathy" were wrong, French assessors judged variations of "mad cow disease" as perfectly right and this caused an important language bias. In our case, five of our subjects thought that "mad cow disease" was a valid answer. If we would have accepted this answer as right, topic 9 would have obtained a global accuracy of 100%.

On the other hand, topics 8, 10, 11 and 16 turned out to be quite easy. Notice that they got an accuracy of 100% in at least one of the proposed systems and they took our subjects fewer time than other topics.

4.5 From passages to documents

We also wanted to analyze the impact of allowing our subject to access the full documents when browsing passages. 29 answers performed with the Passages Searcher was judged as right. In 19 of theses cases, the subject found the answer directly in the passage retrieved by the system, that is, the user wouldn't have needed to visualize the full context. For example, in topic 16 (*When did Edward VIII abdicate?*) the first passage of the ranking contained the answer. In spite of this, most of the subjects used to access the whole document in order to validate the answer and make themselves sure.

On the contrary, when searching topic 8 (*Which airline did the plane hijacked by the GIA belong to?*), the system retrieved passages about GIA's hijackings but it was necessary to check the full context of the paragraph to find out the right answer.

5 Conclusions

In this paper, we have described our participation in the iCLEF 2005 track. We have compared two strategies for finding an answer using an interactive question answering system: i) a search system over full documents and ii) a search system over passages (document's paragraphs). We have added an interesting feature to both system in order to facilitate reading: the possibility to enable/disable the highlighting of named entities such as proper names, temporal references and numbers likely to contain the right answer.

The Document Searcher obtained better overall accuracy (.53 vs. .45) but our subjects found browsing passages simpler and faster. However, most of them presented a similar search behavior (regarding time consumption, confidence in their answers and query refinements) using both systems. Besides, we discuss these data focusing on the causes of failure.

All our users considered helpful the highlighting of named entities. They all extensively used the possibility of emphasize proper names, dates and numbers, specially while the first reading of a long document. They also appreciated the way the Passages Searcher automatically highlighted named entities, according to their initial choices. This feature helped to quickly discriminate between relevant and non relevant passages.

As shown in other CLEF works, it is necessary to count on a good translation of the documents, using MT systems able to distinguish what should and should not be translated. Therefore, we intend to have a more reliable translation of the collections in the future which, without question, will improve the overall results of any cross-language information retrieval experiment.

Acknowledgments

The authors would like to thank the participating subjects. We are also grateful to Valentín Sama and Javier Artiles for their collaboration during the recruitment of volunteers and the achievement of the experiments, and to Anselmo Peñas for his valuable remarks about this paper.

This work has been partially supported by the Spanish Government under project R2D2-Syembra (TIC2003-07158-C04-02). Víctor Peinado holds a PhD grant by UNED (*Universidad Nacional de Educación a Distancia*).

References

[1] J. P. Callan, W. B. Croft, and S. M. Harding. The Inquery Retrieval System. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83. Springer-Verlag, 1992.

[2] C. G. Figuerola, A. F. Zazo, J. L. Alonso Berrocal, and E. Rodríguez Vázquez de Aldana. *Results of the CLEF 2004 Evaluation Campaign*, volume 3491 of *Lecture Notes in Computer Science*, chapter REINA at the iCLEF 2004. Springer Verlag, 2005.

[3] J. Gonzalo and D. W. Oard. *Results of the CLEF 2004 Evaluation Campaign*, volume 3491 of *Lecture Notes in Computer Science*, chapter iCLEF 2004 Track Overview: Interactive Cross-Language Question Answering. Springer Verlag, 2005.

[4] F. López-Ostenero, J. Gonzalo, V. Peinado, and F. Verdejo. *Results of the CLEF 2004 Evaluation Campaign*, volume 3491 of *Lecture Notes in Computer Science*, chapter Interactive

Cross-Language Question Answering: Searching Passages versus Searching Documents, pages 323–333. Springer Verlag, 2005.

[5] V. Peinado, F. López-Ostenero, and J. Gonzalo. UNED at ImageCLEF 2005: Automatically Structured Queries with Named Entities over Metadata. In *Cross Language Evaluation Forum, Working Notes for the CLEF 2005 Workshop*, 2005.