Improving interaction with the user in Cross-Language Question Answering through Relevant Domains and Syntactic Semantic Patterns

Borja Navarro, Lorenza Moreno, Sonia Vázquez, Fernando Llopis, Andrés Montoyo, Migue Ángel Varó

Departamento de Lenguajes y Sistemas Informáticos.
University of Alicante.
Alicante, Spain
{borja,loren,svazquez,llopis,montoyo,mvaro}@dlsi.ua.es

Abstract. The iCLEF 2004 experiment at the University of Alicante has focused on how to assist users in the localization of the correct answer in passages written in a language different from the one of the query. The language of the users is Spanish and the language of the documents/passages English. In order to help users, a first system shows, together with the passage in English, the relevant domains of the passage and the relevant domains of the query. These relevant domains were extracted automatically from WordNet Domains. A second system shows, together with the passage in English, the syntactic-semantic patterns (SSP) of each passage and the SSP of the query. The SSP are formed by the verb and the main nouns of a sentence (that is, the head nouns of the main complements). For users without deep knowledge or with low competence in English, our hypothesis is that to know the relevant domain and/or the SSP will be useful to find the correct answer in the passage. The results show that the SSP are a little bit better in the interaction with the users. However, some users say that it is easier to find the answer knowing the relevant domains than through the SSP.

1 Introduction

The iCLEF 2004 experiment at the University of Alicante has focused on how to assist users in the localization of the correct answer in passages written in a language different from the one of the query. To achieve this objective, we have thought about two important questions:

1. What information must be shown to the user: It must be enough for the efficient localization of the correct answer. The user does not know the correct answer previously. He must infer the correctness of the answer from the context in which it appears. So it is important to show, not only the correct answer, but also enough context that clearly shows that a possible answer is the correct one (or not).

2. How the information is shown to the user: Specifically, in what language the information is shown to the user. If users do not master the language of the passage, they must be helped to identify the correct answer.

In this experiment we have focused on how to assist the user when they do not have enough linguistic competence in the language of the passages. This is the most common case for Spanish people with English language. Most of them know English, but they can not formulate a correct query or understand correctly a possible answer. On other hand, we are looking for alternative methods to deal with large multilingual collection of documents avoiding the use of Machine Translations systems (due to the computational cost of machine translations of the complete collection) [1] [2].

2 Description of the experiment

As we said before, the objective of the experiment is how to assist users in the localization of the correct answer. For this purpose, the experiment has followed the following steps:

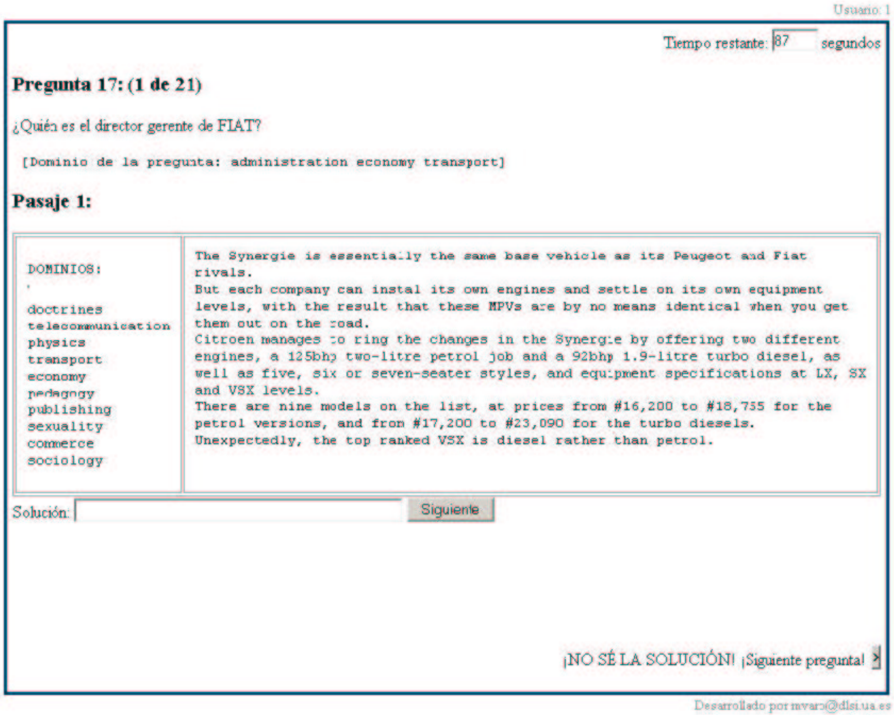
1. **Query formulation and translation.**
   We have taken the queries in Spanish, and they have been translated with a machine translation system to English.

2. **Extraction of relevant passages.**
   For the localization of the relevant passages in the collection of English documents, we have used an Information Retrieval system: IR-n system [3]. This system extracts the passages with a possible answer and ranks them according to probability measures. The size of the passage consist of five sentences, which we think is an optimum size to locate the answer quickly and to infer whether it is correct or not.

3. **Interaction with the users and localization of the answer.**
   The queries (in Spanish) and the passages (in English) are shown to the users through a web page. The users check the passage of each query until they find a passage with a (possible) correct answer. Then they select the answer (the string of characters) and the passage where it appears, and then check the next query.
   The problem is the language: as we said before, the users do not have a deep knowledge of English. They need assistance for the correct localization of the passage and the answer. Ruling out machine translation, two interaction methods have been used in this task. The first one is based on relevant domains [4]: the system shows the users the passage in English, its relevant domains, and the relevant domains of the query. Our hypothesis is that with the relevant domains, the user can decide previously whether or not a correct answer is contained in any passage. The second method is based on syntactic semantic patterns (SSP): the system shows the users the passage

in English and the SSP of each passage, which are formed by the main verbs and the main nouns (that is, the verbs and their subcategorization frame). Our hypothesis is that knowing the SSP, the user can decide whether the passage contains the correct answer. In the next sections, both methods will be explained in depth.

Figure 1 and Figure 2 are the web pages used in the experiment. They show the query, the passage and the relevant domains of the query and the passage (Figure 1); or the query, the passage and the SSP of the passage (Figure 2). If the answer is in this passage, the user selects it, and the system stores the answer, the passage and the time spent. If the answer is not in this passage, the user checks the next passage up to the last one: 50 passages have been extracted from each query.



Tiempo restante: 87 segundos

Pregunta 17: (1 de 21)

¿Quién es el director gerente de FIAT?

[Dominio de la pregunta: administration economy transport]

Pasaje 1:

```
DOMINIOS:
'
doctrines
telecommunication
physics
transport
economy
pedagogy
publishing
sexuality
commerce
sociology
```

```
The Synergie is essentia:ly the same base vehicle as its Peugeot and Fiat
rivals.
But each company can instal its own engines and settle on its own equipment
levels, with the result that these MPVs are by no means identical when you get
them out on the :oad.
Citroen manages :o ring the changes in the Synerg:e by offering two different
engines, a 125bhp two-litre petrol job and a 92bhp 1.9-litre turbo diesel, as
well as five, six or seven-seater styles, and equ:pment specifications at LX, SX
and VSX levels.
There are nine models on the list, at prices from #16,200 to #18,755 for the
petrol versions, and from #17,200 to #23,090 for the turbo diesels.
Unexpectedly, the top ranked VSX is diesel rather than petrol.
```

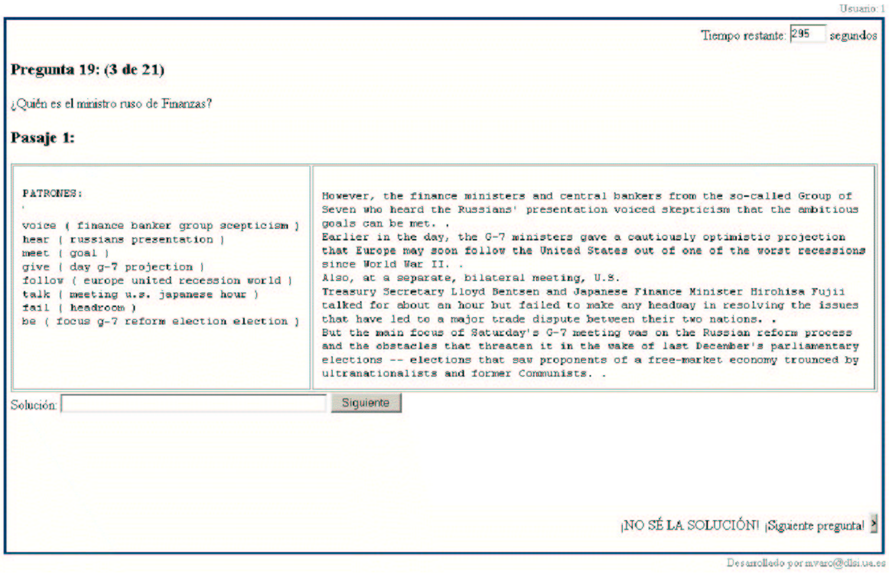
Solución:

Siguiente
¡NO SÉ LA SOLUCIÓN! ¡Siguiente pregunta!
Desarrollado por mvaro@dlsi.ua.es

Fig. 1. Interactive web page with relevant domains.

3 Interaction method I: relevant domains

The fist method uses relevant domains to assist the localization of the correct answer. The relevant domains of a word are the more relevant and representative



Tiempo restante: 295 segundos

Pregunta 19: (3 de 21)

¿Quién es el ministro ruso de Finanzas?

Pasaje 1:

```
PATRONES:
'
voice ( finance banker group scepticism )
hear ( russians presentation )
meet ( goal )
give ( day g-7 projection )
follow ( europe united recession world )
talk ( meeting u.s. japanese hour )
fail ( headroom )
be ( focus g-7 reform election election )
```

```
However, the finance ministers and central bankers from the so-called Group of
Seven who heard the Russians' presentation voiced skepticism that the ambitious
goals can be met. .
Earlier in the day, the G-7 ministers gave a cautiously optimistic projection
that Europe may soon follow the United States out of one of the worst recessions
since World War II. .
Also, at a separate, bilateral meeting, U.S.
Treasury Secretary Lloyd Bentsen and Japanese Finance Minister Hirohisa Fujii
talked for about an hour but failed to make any headway in resolving the issues
that have led to a major trade dispute between their two nations. .
But the main focus of Saturday's G-7 meeting was on the Russian reform process
and the obstacles that threaten it in the wake of last December's parliamentary
elections -- elections that saw proponents of a free-market economy trounced by
ultranationalists and former Communists. .
```

Solución:

Desarrollado por mvaro@dlsi.ua.es

Fig. 2. Interactive web page with SSP.

ontological domains of this word. They are extracted from WordNet Domains (WND) [5]. Our hypothesis is that knowing the relevant domains will help user to decide whether the answer is or is not contained in the passage.

As we said before, this interaction method shows the user the passage in English, its relevant domains, and the relevant domains of the query. Theoretically, the relevant domains of both must agree: the passage with the correct answer must contain the same relevant domains (or very similar relevant domains) as the query. So if users know previously the relevant domains, they can decide whether the answer will or will not be contained in the passage.

WordNet Domains [5] is an extension of WordNet 1.6, where each synset is annotated with one or more domain labels selected from a set of about 250 hundred labels hierarchically organized. To obtain relevant domains, WND glosses are used to collect the more relevant and representative domain labels for each word. Then, the domains associated to the analysed gloss is assigned. All glosses in WN undergo the same process.

We extract the relevant domains of the query and the relevant domains of each passage. This is done through a vector context which represents the relevant domains of the words of the query/passage. From this, we take only the common relevant domains to specify the relevant domains of the whole query/passage.

Furthermore, the passages order has been recalculated according to the similarity between the relevant domains of the query and the relevant domains of the passage. So the system shows first the passage with high similarity between its relevant domains and the relevant domains of the query.

4 Interaction method II: syntactic semantic patterns

The second method is based on syntactic semantic patterns. As we said before, with this method the system shows the user the passages in English and the SSP of each passage, formed by the main verbs and the main nouns (that is, the verbs ant their subcategorization frame). Our hypothesis is that knowing this information, the user can decide whether the passage contains the correct answer and locates it. The intuitive idea is that, when the user is looking for an answer in a text, he focuses on the main nouns and verbs, trying to locate the same or similar nouns/verbs as in the query. With the SSP, the main nouns and verbs have been previously extracted, so maybe they facilitate the task.

From a theoretical point of view, a syntactic semantic pattern is a linguistic pattern formed by three fundamental components [6]:

1. A verb with its sense or senses.
2. The subcategorization frame of the sense.
3. The selectional preferences of each argument.

However, this theoretical SSP is difficult to process automatically: it is difficult to extract patterns like these and to use them in iCL-QA. From this SSP model, we have developed a new one easier to deal with from a computational point of view. In this new model, the verb is represented by the word and its sense (or senses) contained in EuroWordNet; the subcategorization frame is represented by the head noun of each argument[1]; and finally the selectional preferences of each argument are represented by the sense or senses of the head nouns.

With these syntactic semantic patterns, only the most important information of each sentence is shown to the user: the most important words of each sentence –the verb and the subcategorised nouns– and the syntactic and semantic relation between them. As users do not have a deep knowledge of the foreign language (English in our experiment), we think it is better not to process the sentences completely when we are looking for a possible answer. In order to decide whether a passage contains a correct answer, just knowing the main words of the document (that is, the syntactic semantic patterns) will facilitate this task. With these patters, it is difficult to understand completely a text written in a foreign language. However, this is not our objective. Our objective is to find a specific answer for a specific question: first, to decide if the answer is contained in the passage, and then to look for it and find it.

5 Results

5.1 General accuracy

Figure 3 represents the average accuracy obtained by users with each interaction method. This table shows that users achieve similar results with both interaction

[1] If the argument is a clause, the head will be a verb, not a noun. These verbs are, at the same time, a new SSP.

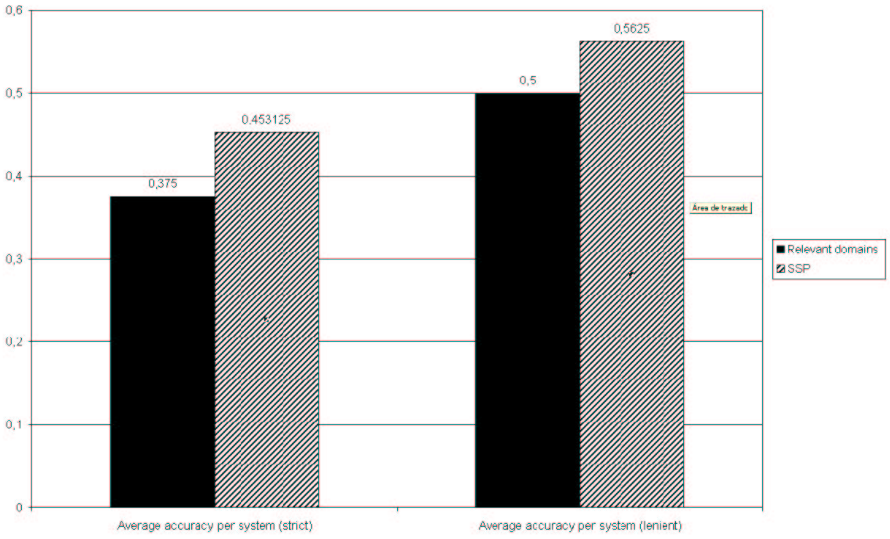


Fig. 3. General average.

methods, but the one based on SSP is a little bit better. From a general point of view, the improvement of the SSP method over the relevant domains method is only 0.015.

5.2 Accuracy by each user

Figure 4 and 5 represent the accuracy achieved by each user. The first table (Figure 4) contains the correct answers located by each user in a passage that really contains the answer ("strict"). In this table, four users locate the correct answer with the correct passage with the SSP method, two users achieve the same results with both methods, and two users achieve better results with the method based on relevant domain.

The second table (Figure 5) shows the correct answers located by each user, independently of the correctness of the passage ("lenient"). In these cases, five users have obtained better results with the interaction method based on SSP, one the same results with both methods, and two users have obtained better results with the relevant domain method.

5.3 Results of the questionnaires

The results of the questionnaire (that users complete during the experiment) do not indicate preferences for any method: five users said that there were no differences between both interaction methods; two users preferred the SSP method, and one the relevant domain method.

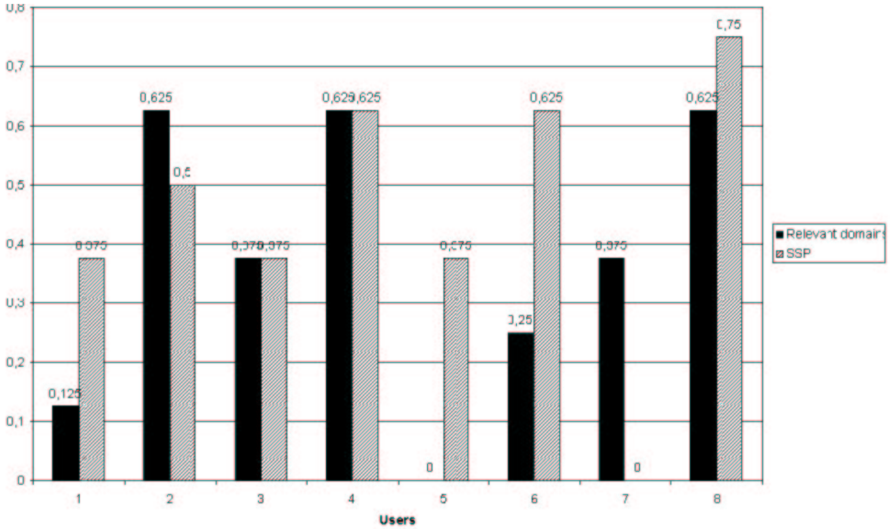


Fig. 4. Strict average user by user.

About the actual help provided by each method (according to the personal opinion of the users), most of them had no preferences. One user clearly preferred the relevant domains method and an other one clearly preferred the SSP method. However, some users said that the actual help of the system in localization the correct answer is rather low in both systems.

5.4 Time consuming

Finally, Figure 6 shows the time consumed by each user. The time consumed with both interactive methods is similar. Two users spent more time with SSP method, and the other six with relevant domains method.

6 Conclusions and future works

From these data, we obtain the next general conclusions:

- The results are low, maybe because we have not used any kind of translation. In this sense, it is necessary some kind of translation (at least, superficial translation) to really help the localization of the answer.

- The order in which the passages are shown to the user in the SSP method (the output order of the IR-n system) seems to be correct.

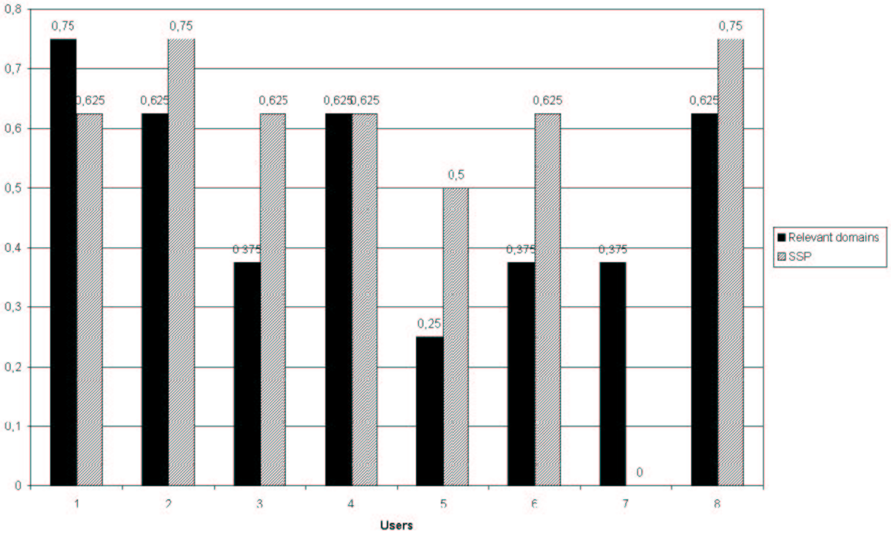


Fig. 5. Lenient average user by user.

- The order in which the passages are shown to the user in the relevant domains method (based on the similarity of relevant domains) does not seem the most correct one.

- According to the results and the personal opinion of the users, relevant domains method really helps the localization of the answer. However, an error in the extraction of the relevant domains will confuse users. For these cases, it is necessary to improve the extraction of relevant domains.

- The SSP method achieve better results, but the users said that it is very difficult to use: it is difficult to read the patterns (only nouns and verb, without any linguistic connection). With this, it is necessary to spend much time reading the patterns. It is necessary to look for an other method to show the patters: for example, to translate the patterns to the language of the user.

- It is necessary to improve the extraction of SSP in order to ensure that the patterns will contain the possible answers: if the answer does not appear in a SSP, the user will not be able to locate it.

The future work of this experiment is focused on two points:

- The syntactic semantic patterns: we are working now in a method to translate the patterns from one language to another based on verb alignment.

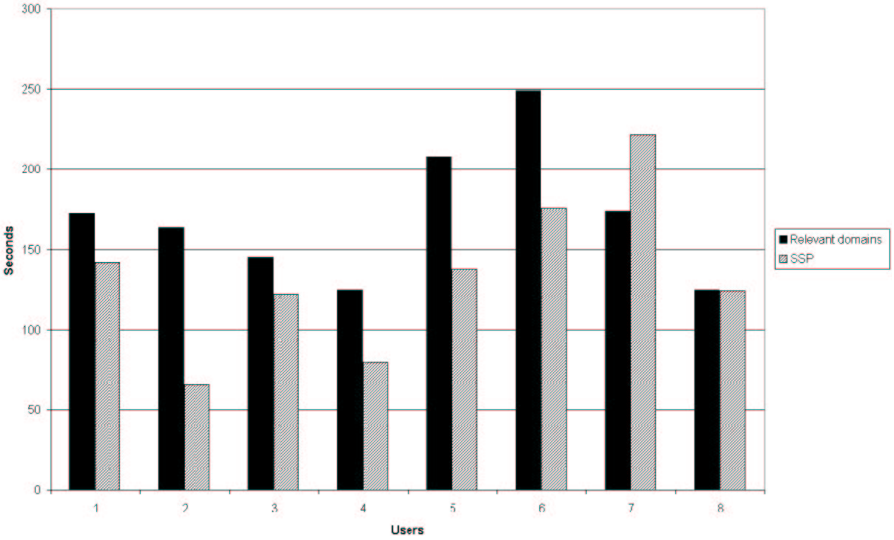


Fig. 6. Time consuming by each user.

- The relevant domains: we are improving their automatic extraction. The idea is to improve the Information Retrieval system with the help of the information on the relevant domains of the passages.

7 Acknowledgements

We want to thank all users for their work and opinions in the development of this experiment.

This work has been partially financed by the Spanish Government (CICYT) with grant TIC2003-07158-C04-01.

References

1. Navarro, B.; Llopis, F. and Varó, MA.: Comparing syntactic semantic patterns and passages in Interactive Cross Language Information Access (iCLEF at University of Alicante). Workshop of Cross-Language Evaluation Forum (CLEF 2003) **Lecture Notes in Computer Science, Springer-Verlang** (2003)
2. López-Ostenero, F.; Gonzalo, J. and Verdejo, F.: UNED at iCLEF 2003: Searching Cross-Language Summaries. Workshop of Cross-Language Evaluation Forum (CLEF 2003) **Lecture Notes in Computer Science, Springer-Verlang** (2003)
3. Llopis, F.: IR-n: Un sistema de recuperación de información basado en pasajes. PhD thesis, University of Alicante (2003)
4. Vázquez S., Montoyo A., Rigau G.: Using Relevant Domains Resource for Word Sense Disambiguation. In: IC-AI'04. International Conference on Artificial Intelligence, Las Vegas (USA) (2004)

5. Magnini, B., Cavaglia, G.: Integrating Subject Field Codes into WordNet. In Gavrilidou, M., Crayannis, G., Markantonatu, S., Piperidis, S., Stainhaouer, G., eds.: Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece (2000) 1413–1418
6. Navarro, B.; Palomar, M. and Martínez-Barco, P.: A General Proposal to Multilingual Information Access based on Syntactic Semantic Patterns. In Anje Düsterhöft and Bernhard Thalheim, ed.: Natural Language Processing and Information Systems - NLDB 2003. Lecture Notes in Informatics, GI-Edition, Bonn (2003) 186–199