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

Borja Navarro, Loren 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 in order to find the correct answer in the passage. The results show that the SSP are few more better in the interaction with the users. However, some users said that it is more easy 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 in 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 do not know the correct answer previously. He/she must infer the correctness of the answer from the context where it appears. So it is important to show, not only the correct answer, but 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 is shown the information to the user. If users are not mastered in the language of the passage, it is necessary to help them in order to identify the correct answer.

In this experiment we have focused on the assistance users when they have low fluency or no 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 it is very common that they can not formulate a correct query or understand correctly a possible answers. On other hand, we are looking for alternative method to deal with large multilingual collection of documents, but avoiding the use of Machine Translations systems (due to the computational cost of the machine translations of the collection completely) [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 propose, the experiment has followed the next 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 measure. The size of the passage is five sentences, that we think it is an optimum size in order to locate the answer quickly and in order to infer if it is a correct answer or not.

3. **Interaction with the users and localization of the answer.**
   The queries (in Spanish) and the passages (in English) are shown users through a web page. The users check the passage of each query until to find a passage with a (possible) correct answer. Then they select the answer (the string of characters) and the passage where it appears, and check the next query.
   The problem is the language: as we said before, the users do not have 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: the system shows user the passage in English and the relevant domains of the passage and the query. Our hypothesis is that with the relevant domains, the user can decide previously if a correct answer is contained in

a passage or not. The second method is based on syntactic semantic patterns (SSP): the system shows user the passage in English and the SSP of each passage, formed by the main verbs and the main nouns of the passage (that is, the verbs and their subcategorization frame). Our hypothesis is that knowing the SSP, the user can decide if the passage contains the correct answer and, finally, he/she can locate it. In the next section, both methods will be explained deeply.

Figure 1 and Figure 2 are the web page 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 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.

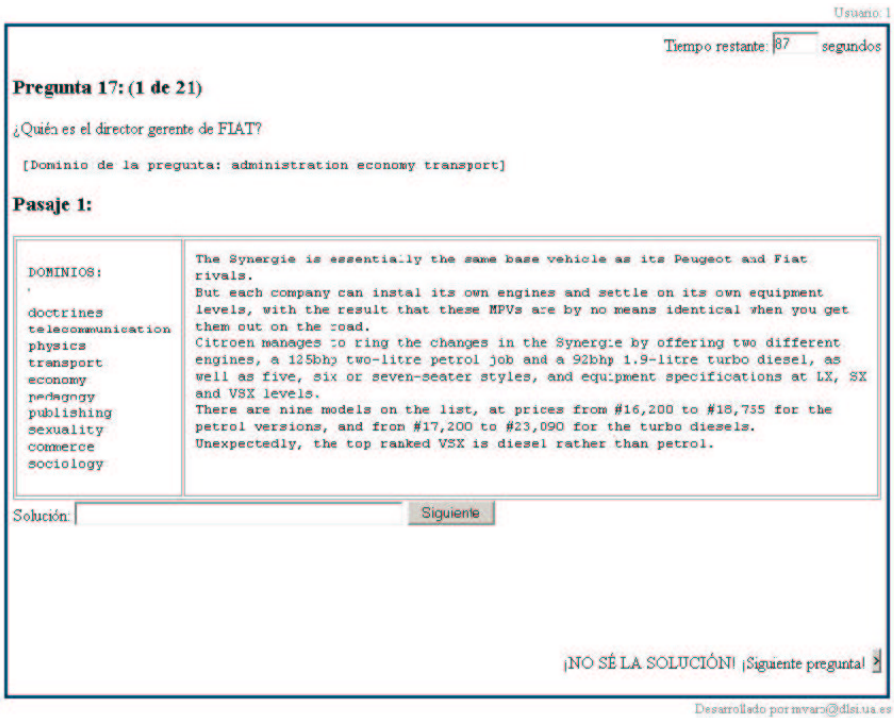

Fig. 1. Interactive web page with relevant domains.

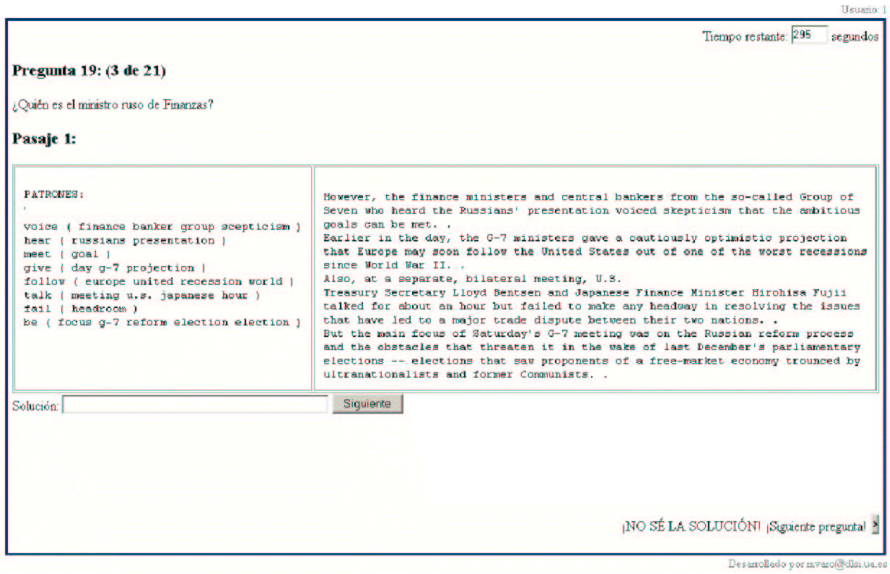

Fig. 2. Interactive web page with SSP.

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 ontological domain of this word. They are extracted from WordNet Domains (WND) [4]. Our hypothesis is that to know the relevant domains will help user to decide if the answer is contained in the passage or not (and then, to look for the answer in the passage).

As we said before, in this interaction method, the system shows user the passage in English and the relevant domains of the passage and 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) than the query. So if users know previously the relevant domains of each one, they can decide previously if the answer will be contained in the passage or not.

WordNet Domains [4] 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 domain associated to the gloss analyzed (each gloss has associated one or more domain labels) is assigned. This process is realized with all glosses in WND. Finally with all this information we can proceed to obtain the relevant domains.

We extract the relevant domains of the query and the relevant domains of each passages. These relevant domains are extracted through a vector context in which the relevant domains of the words of the query/passage are represented. From this, we take only the common relevant domains to specify the relevant domains of the whole query/passage. With this, the relevant domains of the passage or query are the common relevant domain to most of the words.

Furthermore, the passages order has been recalculated according to the similarity between the relevant domains of the 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 query, and at the end the passage with low similarity.

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 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 if the passage contains the correct answer and locate it. The intuitive idea is that, when the user is looking for an answer in a text, he/she looks at the main nouns and verbs, trying to locate the same or similar nouns/verbs than 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 [5]:

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 model of SSP, we have developed a new model more easy to deal with from a computational point of view. In this "lite" model, the verb is represented by the word and its sense (or senses) represented in EuroWordNet; the subcategorization frame is represented by the head noun of each arguments[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 subcategorizated nouns– and the syntactic and semantic relation between them. Due to users have not fluency nor deep knowledge about the foreign language (English in our experiment), we think that it is better not

[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.

to process the sentences completely looking for a possible answer. In order to decide if a passage could contain a correct answer, to know the main words of the document only (that is, the syntactic semantic patterns) will facilitate this task. With this patters, to understand completely a text written in a foreign language is difficult. 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 look for it and find it.

5 Results

5.1 General accuracy

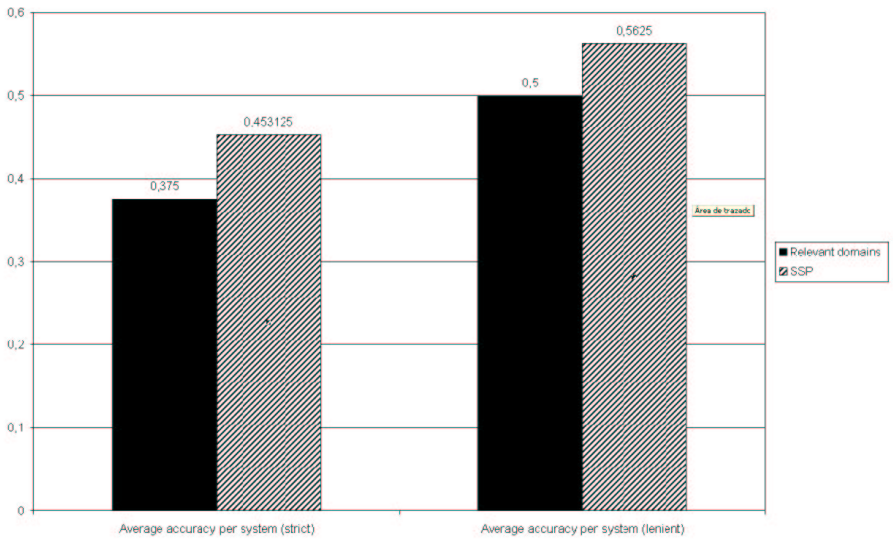

Fig. 3. General average.

Figure 3 represents the average accuracy obtained by users with each interaction method. This table shows that users achieve similar results with both interaction methods, but the one based on SSP is few better. From a general point of view, the improvement of the SSP method from relevant domains method is only 0.015.

5.2 Accuracy user by user

Figure 4 and 5 represent the accuracy achieve user by user. The first table (Figure 4) is the correct answers located by each user in a passage that really contains

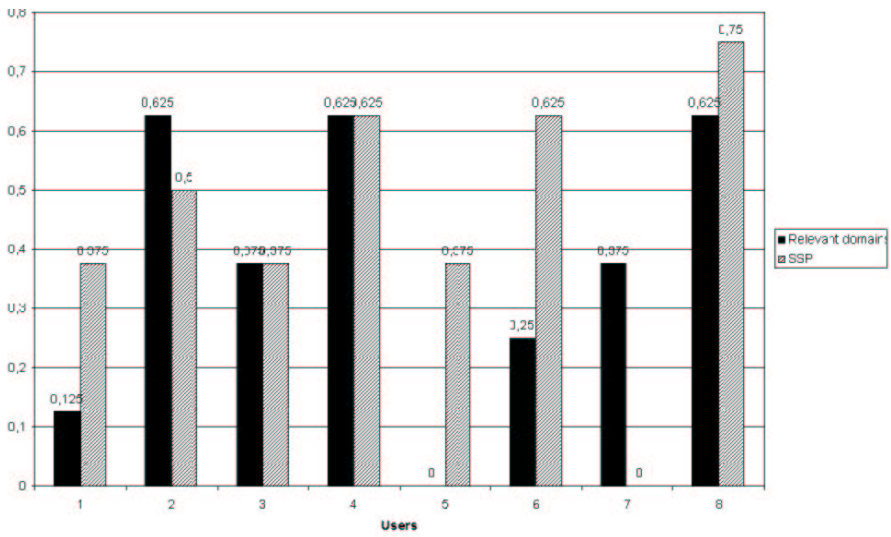

Fig. 4. Strict average user by user.

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 this 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 relevant domain method.

5.3 Results of the questionnaires

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

About the really help of each method (according with the personal opinion of the users), most of them do not prefer one method or other. One user clearly prefers relevant domains method and other user clearly prefers SSP method. However, some users said that the really help of the system in the localization of the answer is low with both systems.

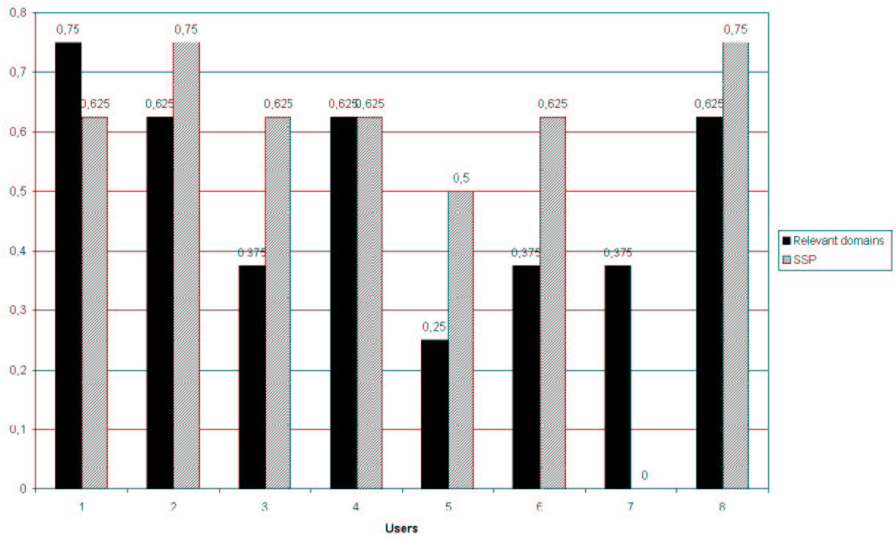

Fig. 5. Lenient average user by user.

5.4 Time consuming

Finally, the Figure 6 shows the time consuming by each user. The time consuming with both interactive method 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 passage are shown to the user in the SSP method (the output order of the IR-n system) seems to be correct.

- The order of passage based on the similarity of relevant domains do not seem the most correct one.

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

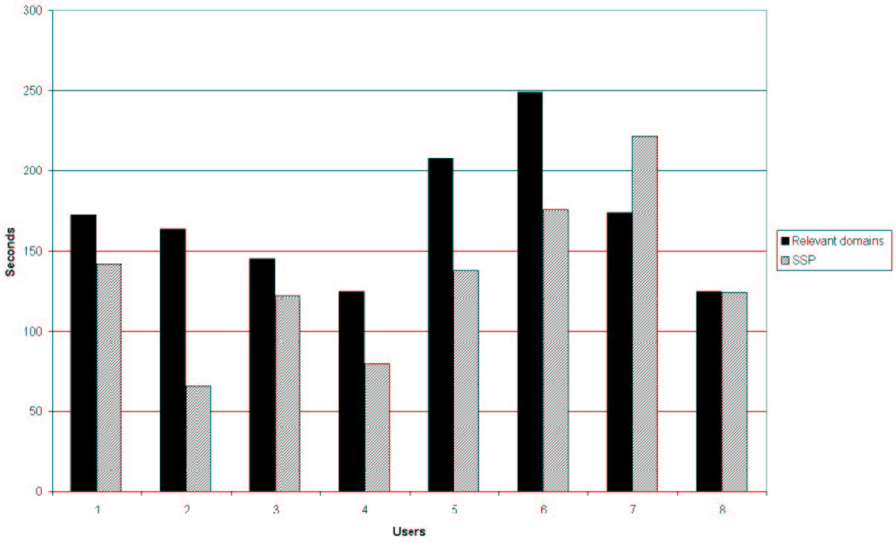

Fig. 6. Time consuming by each user.

- 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 any 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 do not appear in a SSP, the user do not locate it.

The future work of this experiment are focused in two ways:

- About the syntactic semantic patterns, we are working now in a method to translate the patterns from one language to another based on the alignment of the verbs.

- About the relevant domains, we are improving their automatic extraction. The idea is to improve the Information Retrieval system with the information about the relevant domains of the passages.

7 Acknowledgements

We want thanks all users for they work and opinions in the development of the experiment.

This work has been partially supported 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. 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
5. 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