Comparing syntactic semantic patterns and passages in Interactive Cross Language Information Access (iCLEF at University of Alicante)

Borja Navarro, Fernando Llopis, Miguel Ángel Varó
Departamento de Lenguajes y Sistemas Informáticos.
University of Alicante.
Alicante, Spain.
{*borja, llopis, mvaro*}*@dlsi.ua.es*

Abstract

In this paper we will present the result of the interactive CLEF experiment at the University of Alicante. Our aim was to compare two interactive approaches: one based on passages (presented at the iCLEF 2002 [5]), and a new interactive approach based on syntactic semantic patterns. These patterns are composed by the main verb of a sentence plus its arguments, and they are extracted automatically from the passages. With this, these patterns show only the basic information of each sentence. The objective was to know which of these approaches is most useful and fast in the selection of relevant documents by the user in a language different than one of the query (and of the user). The results show that both approach are useful, but the approach based on syntactic semantic patterns is, in the majority of cases, more fast. Finally, with these approaches we avoid the use of Machine Translation systems, due to the problems that they have in Interactive Cross-Language Information Access tasks.

1 Introduction

One of the most important aspects of the Interactive Multilingual Information Access is the way in which the system shows the retrieved documents to the user; mainly, the way in which the system shows the relevant information. Only with this information the user must decide if the retrieved documents are relevant or not. This is a key point in order to ensure the correct selection of documents, and a key point for future refinements of the query.

The main problem is the multilingualism: the user formulates the query in one language, but the relevant documents are written in a different language. To deal with this situation, there are two main solutions: to show the relevant documents to the user in his/her language, or to show the relevant documents in the language of the documents. In the first case, a translation of the document with a Machine Translation system is necessary. However, there are many problems related to Machine Translation. In the second case –to show the information in the language of the document–, it is possible that the user does not be able to understand the information, and does not be able to decide which documents are the relevant ones.

At iCLEF 2002, the University of Alicante proposed the use of passage for the interaction with the user. In this approach, the system selects the most important passage of the retrieved documents. Each passage is translated to the language of the user with a Machine Translation system and, then, the translated passages are shown to the user.

The experiment concluded that this approach based on passages is more fast and has more precision than the approaches based on the whole document. The user only read the relevant passage, not the whole document. This is enough information to decide if the document is relevant or not with high precision. However, there is an important problem with this approach: a lot of passages was unreadable for the user due to problems with the machine translation from English to Spanish [5].

This year we want to improve this approach in two aspects: first, we want to improve the interaction speed –that is, the time consuming by the user between the uploading of the passage to the decision about its relevance–; and second, we want to improve the recall and precision in the selection of relevant documents. On other hand, we want to solve the problem with the Machine Translation systems.

To do this, we have defined an interactive approach based on syntactic semantic patterns [7]. Each syntactic semantic pattern is formed by a verb and the subcategorizated nouns. From a semantic point of view, in each pattern appears the main words of the sentence. We think that it is possible and useful to use these patterns in the interaction with the user, because each pattern contains the main concepts of the document and their syntactic and semantic relations. Instead of showing the passage translated to the user, we show only the syntactic semantic pattern of each sentence in the language of the document (without translation). The users have passive abilities in the foreign language (English), so we think that this is enough information to decide about the relevance of a document.

To conclude, the objectives of our experiment at iCLEF 2003 are:

- to know if it is possible that a searcher decide if a document is relevant or not only with the syntactic semantic patterns extracted automatically from the passage;
- to know if the approach based on syntactic semantic patterns is better than the approach based on passages only;
- to know if the approach based on syntactic semantic patterns is better than the approach based on a machine translation of the passage.

In the next section, we will present these two methods of interaction with the user, the method based on passages and the method based on patterns. Then we will describe briefly the experiment design and the results. Finally, we will show the conclusions and future works.

2 Two methods for the interaction with the users

2.1 Approach based on passages

The first interaction method is based on passage. A passage is the most relevant pieces of text of a document. The main idea of this approach is that it is better to show only the relevant passage to the user, instead of the whole document. This approach was proved at iCLEF 2002 with good results (for more information, see [5]).

2.2 Approach based on syntactic semantic patterns

The second interaction method is based on syntactic-semantic patterns. These patterns are automatically extracted from the passage selected by the Information Retrieval system. The difference between both methods is only how to show the relevant information to the user in a language different from the one of the query: the passage in English only –method one–, or the syntactic semantic patterns extracted from these passages in English too –method two–.

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

1. A verb with its sense or senses.

2. The subcategorization frame of the sense.

3. The selectional preferences of each argument.

For the establishment of this kind of pattern we have take into account several works about subcategorization frame and subcategorization acquisition ([1], [2]), about the relation between verb sense and verb subcategorization ([10], [9]), and about selectional preference ([8], [6]).

For the automatic extraction of these patterns, we have use the syntactic parser Minipar [3]. The extraction system looks for a verb. When a verb is located, it is extracted. Then the system looks for a noun at the left of the verb. If a noun is locates, it is extracted with the verb. Then, the system looks for a noun or preposition plus noun at the right of the verb. If a noun or preposition plus noun is located, it is extracted with the verb and the previous noun. Finally, the system looks for an other noun or preposition plus noun at the right of this noun. If a noun or preposition plus noun is located, it is extracted with the verb and the previous nouns.

For example, from this passage:

"*Primakov suggested that the Administration was using the Ames arrest to score domestic political points, to punish Russia for its independent stance on the conflict in Bosnia-Herzegovina and to provide convenient excuse for cutting American aid to Russia, according to journalists who attended.*"

the system extracts patterns like these:

- Primakov suggest Administration
- administration use Ames arrest
- administration score domestic point
- Primakov punish Russia for its stance
- Primakov provide convenient excuse for
- Primakov cut American aid to Russia according to journalist
- journalist attend

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 the searchers have not fluency nor deep knowledge about the foreign language (English in our experiment), we think that it is better not to process the sentences completely. In order to decide about the relevance of a passage, it is more easy to put the attention on the main words of the document only, that is, to put the attention on the syntactic semantic patterns only.

With this patters, to understand a text written in a foreign language completely is difficult. However, this is not our objective. Our objective is to know the topic of a text or passage and to decide if it is relevant or not.

3 Description of the experiment

We have focused our experiment in the cross-language document selection with a searcher group which has passive language abilities in the foreign language. The language of the user group is Spanish, and the foreign language is English.

The Information Retrieval system used is IR-n system, developed at the University of Alicante [4].The system uses the complete query (tittle, description and narrative) in the search of relevant documents. From each query, the IR-n system locate twenty five (possible) relevant documents.

SYSTEM	F-alpha average
Passages	0.45416703125
Patterns	0.43622984375

Table 1: F-alpha average

Each retrieved document are shown to the user following the order of the experiment design. The first system shows only the passages, and the second system shows the syntactic-semantic patterns extracted from these passages.

Each searcher must decided, showing the passage or the patterns, if the document extracted is relevant or not. As we said before, passages and patterns are written in English, the foreign language. Together with the relevance judgment, we save information about time consuming for each document. Finally, we have developed an HTML interface in order to facilitate the decision of the user about the relevance of each document.

4 Results

The results about the f-alpha average is shown in the Table 1.

These results show that it is possible to decide if a text is relevant or not only with the interaction with the syntactic semantic patterns. The result obtained for each system is very similar (only a difference of 0.0179371875). The first hypothesis is correct.

The time consuming by each searcher during the decision about the relevance of a document is shown in the Table 2. Five of the eight searches consume less time with patterns than with passages. Only in one case, the time consuming with the patterns and with the passages is very similar (searcher 3). Finally, searcher 5 and searcher 6 use much more time with the patters than with the passage. However, this is an abnormal time consuming, because the different time consuming between passages and patterns is very large. Probably, due to problems during the experiment. Finally, the searcher that has obtain the better result (searcher 4) consumed more time with the passages than with the patterns.

With this data, we can conclude that the use of pattern in the interaction process improve the time consuming in the majority of cases. So the second hypothesis is, in some way, correct.

Finally, with the syntactic semantic pattern it is possible to avoid the use of Machine Translation systems. The results obtained this year at the iCLEF 2003 are better than the results obtained at iCLEF 2002, in which a Machine Translation system was used.

5 Conclusions

In this experiment, we have compare two method of interaction with a IR system, the first one based on passages and the second one based on syntactic semantic patterns. The results show that it is possible to decide if a document is relevant or not with the syntactic semantic patterns only. On other hand, the time consuming is less with the patterns than with the passages in the majority of cases. Finally, with these patterns it is possible to avoid the use of Machine Translation systems.

These syntactic semantic patterns are a simplification of the language: each pattern contains the main concepts and linguistic relations of a sentence. Due to this simplification, it is possible the use of these patterns in other cross-linguistic information access task as, for example, the indexation and search of documents by patterns, the alignment of pattern extracted from different languages (through the verb), or the refinement of the query with the patterns contained in the documents selected by the user.

System	Searcher 1	Searcher 2	Searcher 3	Searcher 4
Passages	4359	5641	5361	5533
Patterns	3195	4840	5548	3063

System	Searcher 5	Searcher 6	Searcher 7	Searcher 8
Passages	1350	1707	1835	5287
Patterns	829	5046	7957	2555

Table 2: Time consuming

6 Acknowledgements

We would like to thank the users (Belén, Raquel, Irene, Julio, Rafa, Ángel, Sonia and Yenori) and Rubén, who implemented the pattern extraction system.

References

[1] Ted Bricoe and John Carroll. Automatic Extraction of Subcategorization from Corpora. In *Workshop on Very Large Corpora*, Copenhagen. 1997.

[2] Anna Korhonen. *Subcategorization acquisition.* Technical Report. University of Cambridge, Cambridge, 2002.

[3] Dekang Lin. Dependency-based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*, Granada. 1998.

[4] Fernando Llopis. *IR-n: Un sistema de recuperación de información basado en pasajes.* PhD thesis, University of Alicante, 2003.

[5] Fernando Llopis, Antonio Ferrández, José Luis Vicedo, Manuel Díaz, and Fernando Martínez. iCLEF at Unversities of Alicante and Jaen. *Workshop of Cross-Language Evaluation Forum (CLEF 2002)*, Lecture Notes in Computer Science, Springer-Verlang, 2002.

[6] Diana McCarthy. *Lexical Acquisiton at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences.* PhD thesis, University of Sussex, 2001.

[7] Borja Navarro, Manuel Palomar, and Patricio Martínez-Barco. A General Proposal to Multilingual Information Access based on Syntactic Semantic Patterns. In Anje Düsterhöft and Bernhard Thalheim, editor, *Natural Language Processing and Information Systems - NLDB 2003*, pages 186–199. Lecture Notes in Informatics, GI-Edition, Bonn, 2003.

[8] Philip S. Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships.* PhD thesis, University of Pennsylvania, 1993.

[9] Douglas Roland. *Verb Sense and Verb Subcategorization Probabilities.* PhD thesis, University of Colorado, Colorado, 2001.

[10] Douglas Roland and Daniel Jurafsky. Verb sense and verb subcategorization probabilities. In P. Merlo and S. Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimetal Issues*, pages 325 – 346. John Benjamins, Amsterdam, 2002.