A general proposal to Multilingual Information Access based on syntactic-semantic patterns.

Borja Navarro, Manuel Palomar, and Patricio Martínez-Barco.
Departamento de Lenguajes y Sistemas Informáticos.
Universidad de Alicante.
Alicante, Spain
{borja,mpalomar,patricio}@dlsi.ua.es

Abstract:

In the last few years, there has been a wide development in the research on textual information systems. The goal is to improve these systems in order to allow an easy localization, treatment and access to the information stored in digital format (Digital Databases, Documental Databases, and so on). There are lots of applications focused on information access (for example, Web-search systems like Google or Altavista). However, these applications have problems when they must access to cross-language information, or when they need to show information in a language different from the one of the query.

This paper explores the use of syntactic-sematic patterns as a method to access to multilingual information, and revise, in the case of Information Retrieval, where it is possible and useful to employ patterns when it comes to the multilingual and interactive aspects. On the one hand, the multilingual aspects that are going to be studied are the ones related to the access to documents in different languages from the one of the query, as well as the automatic translation of the document, i.e. a machine translation system based on patterns. On the other hand, this paper is going to go deep into the interactive aspects related to the reformulation of a query based on the syntactic-semantic pattern of the request.

1 Introduction.

In the last few years, there has been a wide development in the research of textual information systems. The goal is to improve these systems to allow an easy localization, treatment and access to the information stored in digital format (Digital Databases, Documental Databases, and so on).

At present, there are lots of applications focused on information retrieval. The main ones are Web-search systems (like Google or Altavista). However, these applications are not totally satisfactory due to the complexity and multilinguism of Internet.

The “Multilingual Information Access” is a new step in the research in treatment and access to textual information. It is centered basically on Internet, and its main characteristic is the treatment of information from heterogeneous and multilinguistic sources, although not from documents. The Multilingual Information Access is a large research area that

includes subareas like “Multilingual Information Retrieval”, “Interactive Multilingual Retrieval”, “Multimedia Information Access”, “Question Answering”, and so on.

One of the main problems in Multilingual Information Access is the treatment of cross-language information, that is, how to deal with information which is encoded in languages different from the one of the query. In fact, the query must be “translated” (in some way) into the languages in which the information is encoded and this information must then be “translated” (in some way) into the language of the user.

To deal with this problem, we propose in this paper the use of syntactic-semantic patterns as a method of accessing multilingual information. In the case of Multilingual Information Retrieval, we explore and revise where it is possible and useful to employ patterns when it comes to the multilingual and interactive aspects. On the one hand, the multilingual aspects that are going to be studied are the ones related to the access to documents in languages different from the one of the query, as well as the automatic translation of the document, i.e. a machine translation system based on patterns. On the other hand, the interactive aspects which the paper is going to explore are the ones related to the reformulation of the query based on the syntactic-semantic pattern of the request.

Theoretically, a pattern is basically a linguistic pattern formed by three fundamental components:

1. A verb with its sense or senses.
2. The subcategorization frame of the sense (that is, the syntactic category of each argument of the verb).
3. The selectional preferences of each argument (that is, the main semantic features of each argument).

For the establishment of this kind of pattern we have taken into account several works on subcategorization frame and subcategorization acquisition ([Bri97], [Kor02]), on the relation between verb sense and verb subcategorization ([Rol02], [Rol01]), and on selectional preferences ([Res93], [McC01]).

However, a linguistic pattern in a Multilingual Information framework must be supra-linguistic or cross-linguistic, therefore, together with these three components, a fourth component is added: a multilingual component that allows the use of this pattern in Multilingual Information Extraction. This multilingual component is an identifier of alignment between two or more patterns in different languages. These aligned patterns share the same components (the same verb sense, similar arguments and similar selectional preferences), but in different languages.

The Multilingual Information Access is a relatively modern task. The first evaluation campaign on it was undertaken in 1996. There are currently several approaches to the problem of multilingual aspects: the most common one is the translation of the query to the languages of the documents. Dictionaries, thesauri, parallel and comparable corpora, and Machine Translation Systems are some of the resources used in this task. Another approach is the translation of the documents. Due to the large size of the collections, normally “shallow”

translations are used. Finally, there are other approaches like conceptual indexation with EuroWordNet, that has shown good results [Lóp02].

In a whole Multilingual Information Retrieval the interaction with the user is necessary for him to refine the query and make it more precise in locating his information request. The main method to interact with the user is the "relevance feedback", where the user selects the documents more related with his information request, and the system analyzes them and reformulates the query with the information extracted from the selected documents [Hea99], [Lóp02].

2 A proposal to Multilingual Information Access based on syntactic-semantic patterns.

From a general point of view, the Multilingual Information Access task consists in searching an idea or concept (which is contained in a query) from a large amount of multilingual documents. To process the query and the documents in order to achieve this goal, there are different approaches such as methods based on keywords, methods based on noun phrases ([Lóp02]) and others (see [Pet02]). All these methods try to solve the following problems: 1. The extraction of the idea or concept from the query, 2. The treatment of large amounts of documents, 3. The multilingual nature of the collection of documents, 4. The interactivity with the user in order to refine the query.

In our case, we define a method to information access based on the use of syntactic-semantic patterns, as we have defined in the previous section. These four problems can be specified in six subtasks: 1. The extraction of a pattern from the query, 2. The extraction of patterns from a large amount of multilingual documents,3. The indexation of the collection of documents based on patterns from each language, 4. The search of relevant documents from the multilingual collection, 5. The presentation of a summary from the relevant documents for the user, 6. The interaction with the user aimed at refining the query.

In the next sections, we will develop each one of the tasks and the use of patterns in them.

2.1 Extraction of the head pattern from the query.

In the iCLEF forum[1], there are two kinds of queries: broad queries which ask about several aspects of the topic, and narrow queries which ask about a specific aspect of the topic. For example, this is a narrow query used in the iCLEF 2001 [Oar02].

NUMBER: C029

TITLE: Nobel Prize for Economics

[1] http://terral.lsi.uned.es/iCLEF/

DESCRIPTION: Who won the Nobel Prize for Economics in 1994 and for what theory?

NARRATIVE: Find documents that give both the names of the winners of the Nobel prize for Economics in 1994 and the name of the theory for which the prize was awarded.

The queries used in this forum (called "topics") are extracted from the CLEF 2000 forum[2]. They are information requests formulated by hypothetical users, structured in three files: title, description and narrative. The title is a noun phrase with no more than three words. The description is a long sentence (normally, in the narrow topics). Finally, the narrative is a more elaborated formulation of the topic with several sentences [Bra02].

Each participant in the CLEF 2000 forum and in iCLEF 2001 used the fields of the query they considered important for their objectives. The majority of experiments used the title and the descriptions, others used the three fields, and still others the title field only, and so on.

The Multilingual Information Access based on patterns that we propose is focused on the fields of the topic formed by sentences (the description mainly, but the narrative too) , because the sentence is necessary to extract a pattern. From the sentence, we extract the pattern, where the main words and the relation between them is expressed.

For example, in the above description, the following sentence appears:

Who won the Nobel Prize for Economics in 1994 and for what theory?

The pattern extracted from the sentence will be:

Win - Nobel Prize

In the pattern extracted (which we call "Head Pattern"), we show the relation between the verb and the object. However, we do not use this pattern to search in the collection of documents directly, but generate a structure formed by specific patterns instead, which are derived from the head pattern. The derivation is made through syntactic and semantic information. We call this derived patterns "Syntactic-Semantic Patterns".

From the syntactic point of view, the head pattern is expanded by all the verb forms. From the semantic point of view, the head pattern is expanded with the semantic relation contained in the WordNet of each language [Vos99] for each verb extracted. The semantic relation we use are: synonymy, hyponymy, hyperonym, "cause" and "is-caused-by", "rol" and "involved", "co-role", "has-subevent" and "is-subevent-of".

For example, in Figure 1 we show the derived structure of the head pattern "win - Nobel Prize". Thus, the patterns derived are:

Wins Nobel Prize; Take Nobel Prize; Taking Nobel Prize; Winning Nobel Prize; Overcomes Nobel Prize; Contended Nobel Prize; Thrash Nobel Prize; ...

[2] http://clef.iei.pi.cnr.it/

With this method we may come up with unreal patterns. For example, in the above expanded patterns we find the syntactic-semantic pattern “Thrash Nobel-prize”, but as there are no documents with this pattern, no documents will be extracted.

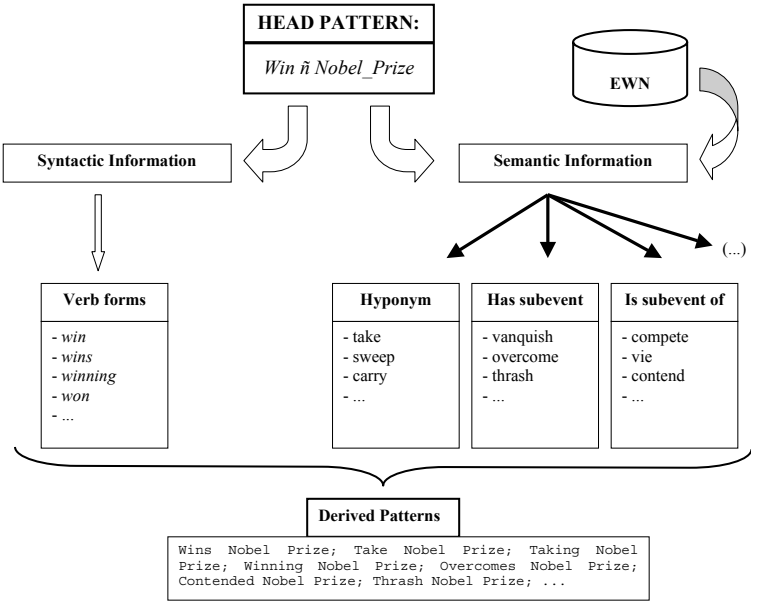

Abbildung 1: Expansion of the head pattern: the syntactic-semantic pattern structure.

2.2 Extraction of the patterns from a collection of documents.

To deal with the collection of documents through syntactic-semantic patterns, we extract all the head patterns. The mechanism of extraction is the same for the extraction of the head pattern from the query. This will be explained later.

This way, the treatment of the collection of documents is more complete, because the search is done through the syntactic-semantic patterns, that is, through all the patterns derived from the head pattern and the patterns derived from the collection of documents, and not through the words of the documents or the topic.

2.3 Indexation of the collection of documents by patterns.

Once the head patterns are extracted from the collection of documents, they are indexed. Usually, the indexation is made by words. However, we propose an indexation by head patterns where each head pattern is specifically linked with the document in which it ap-

pears.

Due to the multilingual nature of the collection of documents, the indexation must be made in each language independently. We obtain a different indexation from each language. In other words, each document is indexed in its own language.

2.4 Monolingual and multilingual search for relevant documents.

In the process of searching, we first search the patterns derived from the query among the patterns extracted from the documents which are in the same language as that of the query. A document is considered relevant if it contains a syntactic-semantic pattern derived from the query and/or if it contains a pattern with semantic similarity to the head patterns. This semantic similarity is based on the semantic distance of the words of the patterns in the WordNet of each language [Bud01].

However, to deal with the patterns extracted from documents written in other languages, we need a multilingual search: to do this, we use a database of aligned patterns.

As we will explain later, the database of aligned patterns is extracted from comparable corpora (that is, a corpora formed by texts from different languages -usually two-, dealing with the same or similar topics). The patterns extracted from each language are aligned through the Interlingua Index of EuroWordNet: two patterns are aligned if they share the same verb sense and if they have a similar sense in their arguments.

With this database, a head pattern in one language is expanded by the patterns related in other languages (see Fig. 2). These patterns are used to search in the indexation of the documents of each language. Just as we have done for monolingual search, in the multilingual one we calculate the semantic similarity between the syntactic-semantic patterns and the patterns indexed in the documents. In this case, due to the multilingual nature of the documents, we calculate a multilingual semantic similarity. Thus, we select the documents related to the syntactic-semantic pattern in each language, which are considered relevant.

2.5 The presentation of the selected documents and the interaction with the user.

Once the relevant documents are extracted, they are displayed to the user. However, at this step there are three aspects to be looked into:

1. The summarization: The documents selected could be too many. It is better to show just a summary containing the most important information of the document.

   Our proposal is to show a list of the relevant patterns extracted from each document. This list constitute a brief summary of the document, because the patterns hold the information of the documents related to the query.

   Let's take, for example, the following text extracted from the CLEF collection:

   THOMAS J. WATSON JR.; LED IBM INTO COMPUTER AGE. Thomas J. Watson

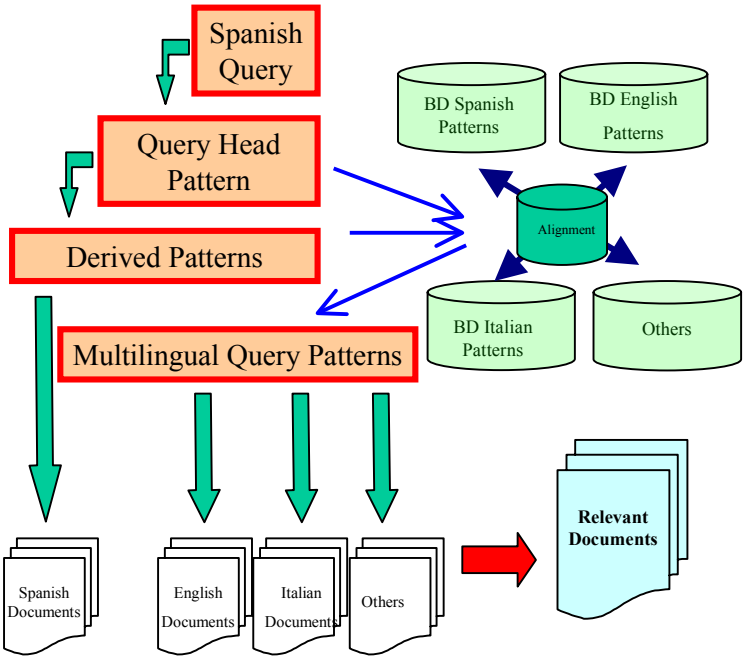

Abbildung 2: Multilingual Expansion of the head pattern

Jr., who brought International Business Machines – the firm founded by his father – from the age of typewriters into the era of computers, died Friday. A company spokesman said Watson, who spent years feuding with his father over the direction IBM should take, was 79 when he died at Greenwich Hospital in Greenwich, Conn., after a stroke. He had retired in 1971 after a heart attack and later served as ambassador to the Soviet Union.

According to the patterns extraction which we will show later, the patterns extracted are similar to the following:

1. *lead - IBM - Computer age (into)*; 2. *who - bring - International Business Machines*; 3. *firm - found - father (by)*: 4. *die - Friday*; 5. *spokesman - say*; 6. *who - spend - years*; 7. *feud - father (with) - direction (over)*; 8. *IBM - take*; 9. *be - 79*; 10. *he - die - Greenwich Hospital (at) - Greenwich (in)*; 11. *he - retire - 1971 (in) - heart attack (after)*; 12. *serve - ambassador (as) - Soviet Union (to).*

With just this list of patterns, it is possible to know the topic of the text and what the author is talking about. Thus the list of patterns acts as a summary of the text.

2. Multilinguism: the summary must be came in the language of the query, that is, in the language of the user. However, due to the multilingual nature of the collection,

the relevant documents could be written in a language different from that of the topic.

To solve this problem, we use the database of aligned patterns. Instead of showing the list of patterns in the language of the document (the original language), we show the list of aligned patterns in the language of the user (Fig. 3).

3. Interaction with the user and refinement of the query: The relevant documents could be too general and not relevant to the user.

   To solve this problem, we can refine the query by introducing new, more specific information in the head pattern, such as information about time or places; that is, by the introduction of adjunct complements.

   For example, to the above pattern

   Win - Nobel Prize

   the document selected are concern Nobel Prize winning. However, it can be refined with two adjuncts contained in the description: information about the year and the category of the prize. The remaining pattern will then be:

   Win - Nobel Prize (Economics, 1994)

   The information request is now more specific, as will be the extracted documents.

   Unlike the first search of patterns, these refined patterns are not searched in the whole collection of documents, but in the relevant documents previously selected, which are more specific and closer to the requirements of the user. The objective is to locate the most relevant documents for the user from the set of selected documents.

   This refinement of the query is done necessarily by interacting with the user. An effective technique is the “relevance feedback”, which allows the user to evaluate the documents selected in the first search and to indicate the relevant ones for his request of information. The system analyzes these documents and extracts new information to refine the query [Lóp02].

   When the process of information retrieval is done by syntactic-semantic patterns, the refinement of the query is focused on the extraction of more information which will enrich previous patterns and/or generate new ones with more specific information. The objective in the analysis of the documents selected by the user is to extract the patterns of each document and select the patterns common to all documents (together with the patterns common to the original one). These patterns contain more precise information and will be come the new head patterns in a new search for information.

   The patterns extracted from the documents selected must be displayed to the user, who then has the possibility of rejecting those which do not adjust to his request, and give priority to those which do adjust to his request.

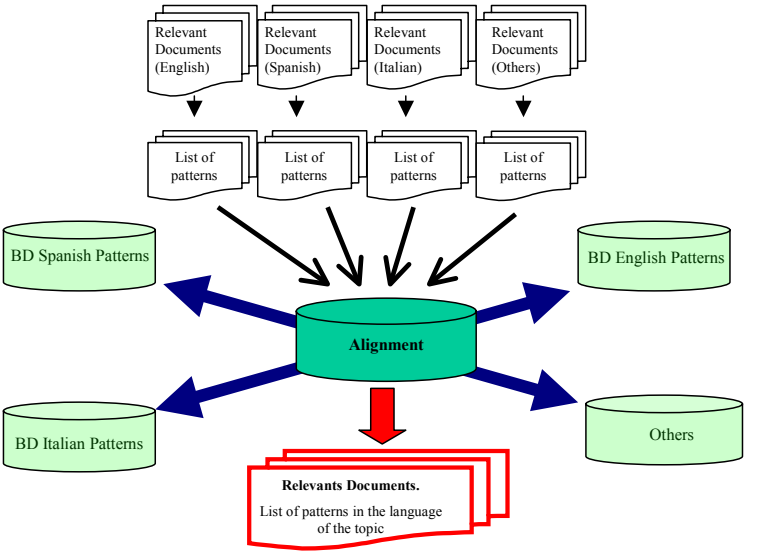

Abbildung 3: Showing the list of patterns in user language

For the objectives of this general proposal there are two main problems that must be treated previously: 1. The extraction of head patterns, and the generation of the structure of syntactic-semantic patterns, 2. The alignment of patterns between different languages.

In the next section we will explain the method used to extract patterns, and finally we propose an algorithm for aligned patterns in an interlingua framework.

3 Automatic method for the extraction of patterns.

With the method of extracting head patterns, we have two general objectives:

1. The extraction of head patterns from the “topic query” and the extraction of head patterns from the collection of documents.
2. The extraction of pattern from a comparable corpora in order to align them and create a database of aligned patterns.

The method of extraction is the same for both tasks.

To do this, it is necessary to use several Natural Language Processing (NLP) tools, because, as we have shown above, the patterns are based on linguistic information. A pattern contains a verb, some complements (the arguments of the verb), and the relation between

them. In order to know this, we need the part of speech of each word and the general structure of the sentence. Besides, in order to know the semantic characteristics of each component, we need to know the sense (or senses) of each word and the semantic relations between them.

The NLP tools we need are the following: a PoS-tagger, a syntactic parser and a Word Sense Disambiguation system (WSD).

This is an ideal situation. However, the application of NLP tools presents several problems. For example: the application of NLP tools such as syntactic parsers to large collections of documents uses too much computational time and resources; and, on other hand, the use of WSD systems generates some errors, because the word sense disambiguation is still an open problem. The better systems achieve 70% accuracy in English.

For these reasons, in this first proposal we have limited the use of NLP tools.

First of all, we use a PoS-tagger to obtain the lemma of the words and their part of speech, but, instead of using a syntactic parser, we have define a regular expression that represents the basic syntactic relations between a verb and the nouns. A head pattern is the set of expressions derived from the following regular expression:

Head Pattern:
`[noun]verb[noun|preposition_noun][noun|preposition_noun]`

Thus, we avoid the use of a robust syntactic parser and the problems it represents.

The general mechanism for obtaining the patterns is as follows: we look into a lemmatizated and PoS annotated corpus for verb-noun relations that coincide with the predefined regular expression. When an expression is located, a head pattern is generated.

For example, from this fragment:

Watson falleció en el hospital Greenwich a consecuencia de complicaciones tras sufrir un ataque cardíaco, añadió la fuente. El difunto heredó de su padre una empresa dedicada principalmente a la fabricación de máquinas de escribir y la transformó en una compañía líder e innovadora en el mercado de las computadoras. (...) China sustituyó hoy, sábado, su doble sistema de cambio monetario y fijó su primera y única tasa flotante para el yuan, que pasó a tener un valor de 8,7 por dólar, informó la agencia de prensa Nueva China. La tasa de cambio oficial china era hasta ayer de 5,8 yuanes por dólar para residentes extranjeros y turistas, y una tasa flotante fijada en función de la oferta y la demanda, sobre el mercado Swap, para las empresas. EFE

we extract the following head patterns: 1. *fallecer - hospital*; 2. *sufrir - ataque cardíaco*; 3. *añadir - fuente*; 4. *difunto - heredar - padre (de) - empresa*; 5. *dedicar - fabricación (a) - máquinas (de)*; 6. *transformar - compañía (en)* (...) 7. *China - sustituir - sistema - cambio monetario (de)*; 8. *fijar - tasa - yuan (para)*; 9. *tener - valor - 8'7 (de)*; 10. *informar - agencia - prensa (de)*; 11. *cambio - era - 5'8 (de) - yuanes*; 12. *tasa - fijar - función (en) - oferta.*

In this example we have found a few problems:

- Multiwords: The head of the noun phrase could be a compound noun (for example, "ataque cardíaco"[3]) and not a simple noun ("ataque"[4]), so some strategy is necessary to locate these complex heads or multiwords.
- Loosing information: In some patterns we lose a lot of information because we only extract the head of the phrases situated in a position close to the verb. However, a phrase with relevant information might be situated in a position distant from the verb. For example, in the phrase "máquina de escribir"[5], where only the single word "máquina"[6] has been extracted in the pattern.
- Named Entity: proper names are not stored in WordNet, therefore we need a Named Entity Recognition to locate them.
- Free order of constituents: In the third pattern, for example, the noun "fuente"[7] appears on the right position of the verb, which is the typical position of the object, although, it has the function of subject. This is a general problem in the free order of Spanish and other romance languages.

To solve this problem, we need to use a parser, but, as we said before, the use of syntactic parsers with large collections of documents is not satisfactory. We will later make some experiments with syntactic parsers in order to find the strategy which best allows us to extract the correct patterns with the least computational time and resources.

In this first proposal we do not use Word Sense Disambiguation systems, due to their high degree of errors. Instead of this, we take all the senses of the words which have more than one sense (ambiguous words). Therefore, we do not resolve the ambiguity. However, in next experiments we will use these systems to see if there is any improvement in the extraction of the patterns.

4 A database of aligned patterns.

Finally, in this paper we present a proposal for the alignment of patterns in different languages, in order to create a database of aligned patterns.

To search in corpora of different languages, we need a methodology that allows us to cross from a text written in one language to one written in another language. To do this with the patterns, we propose the use of a cross-lingual alignment.

From a general point of view, two patterns of different languages align if they have the same verb sense and if they have similar senses in their arguments. At a conceptual level, these two patterns are the same, because they express the same idea or the same concept.

[3] *Heart attack*
[4] *Attack*
[5] *Typewriter*
[6] *Machine*
[7] *Source*

The only difference is that each one of them is written in a different language. This is the reason why it is possible to establish a cross-lingua link between them.

The alignment of patterns is based on:

- The use of the Inter Lingual Index of EuroWordnet as an index of alignment: this is the reference to establish a link between two patterns. Two patterns align only if their verbs are aligned in the Inter Lingual Index of EuroWordNet.
- The use of comparable corpora as a training corpus from where the patterns can be extracted. A comparable corpora is formed by texts from different languages – normally two– with the same or similar topics. Thus, the patterns extracted from the texts of each language will have the same topics and a high probability of alignment between them.

The alignment of patterns is based on the following characteristics. Two patterns are aligned if they satisfy two rules:

- First: their verbs are aligned in the Internligual Index of EuroWordNet.
- Second: their complements are similar.

  We consider two complements as similar in four cases:

  – first, if they are aligned in the Interlingual Index;
  – second, if their hyperonyms are aligned in the Interlingual Index;
  – third, if one of them is aligned with the hyperonym of the other one in the Interlingual Index;
  – fourth, if the hyperonym of one of them is aligned with the other one in the Interlingual Index.

For example, the two previous texts have the same topic: the death of Thomas J. Watson, manager of IBM. They are not parallel texts, so there is not a complete alignment between all the sentences of the texts. However, due to the fact that they are talking about the same topic, it is possible that some sentences (probably, the main ones) align between them.

This is the list of patterns extracted from the English text:

1. *lead - IBM - Computer age (into)*; 2. *who - bring - International Business Machines*; 3. *firm - found - father (by)*; 4. *die - Friday*; 5. *spokesman - say*; 6. *who - spend - years*; 7. *feud - father (with) - direction (over)*; 8. *IBM - take*; 9. *be - 79*; 10. *he - die - Greenwich Hospital (at) - Greenwich (in)*; 11. *he - retire - 1971 (in) - heart attack (after)*; 12. *serv - ambassador (as) - Soviet Union (to).*

And this is the list of patterns extracted from the Spanish text:

1. *fallecer - hospital*; 2. *sufrir - ataque cardíaco*; 3. *añadir - fuente*; 4. *difunto - heredar - padre (de) - empresa*; 5. *dedicar - fabricación (a) - máquinas (de)*; 6. *transformar - compañía (en).*

Through the similarity of the verb sense, the pattern

"Fallecer - hospital"

aligns with the pattern

"He - die - Greenwich Hospital (at) - Greenwich (in)"

These two patterns have the same verb sentence (*to die*) and similar sentences in their arguments (*hospital*). With this technique, a large comparable corpora is processed.

Finally, the head patterns extracted and aligned are stored in a database, which acts as a multilingual dictionary for the search of patterns in a multilingual collection of documents.

5 Conclusions and Future works.

In this paper we have shown a general proposal for the use of syntactic-semantic patterns in Multilingual Information Access. First we have described the patterns from a linguistic point of view. Then we have shown the general process of using these patterns in the Multilingual Information Access. To do this, we have proposed a method for extracting head patterns from the corpora and a method for aligning them, in order to build a database of aligned patterns.

Future works in this general proposal will focus on its improvement. Several NLP tools will be tested, especially, the use of syntactic parsers and the use of a Word Sense Disambiguation systems.

Our objective is to participate in the next iCLEF 2003 forum, in order to evaluate the multilingual and interactive aspects of our general proposal. [8]

Literatur

[Bra02] Braschler, Martin. CLEF 2001 - Overview of Results. In Peters, Carol; Braschler, Martin; Gonzalo, Julio and Kluck, Michael, editor, *Evaluation of Cross-Language Information Retrieval Systems*, pages 9 – 26. Lecture Notes in Computer Science. Springer, New York, 2002.

[Bri97] Bricoe, Ted and John Carroll. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5 th ANLP Conference. ACL*, pages 356– 363, Washington D.C., 1997.

[Bud01] Budanitsky, Alexander and Graeme Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001. North American Chapter of the Association for Computational Linguistics (NAACL).

[8] The developments, improvements and prototypes of this general proposal will appear in the proposal web site: `http://www.dlsi.ua.es/ borja/patterns.html`.

[Hea99] Hearst, Marti A. User Interfaces and Visualization. In Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors, *Moder Information Retrieval*, pages 257–340. Addison-Wesley, ACM Press, New York, 1999.

[Kor02] Korhonen, Anna. *Subcategorization acquisition*. Technical Report. University of Cambridge, Cambridge, 2002.

[Lóp02] López Ostenero, Fernando. *Un sistema interactivo para la búsqueda de informatión en idiomas desconocidos por el usuario*. PhD thesis, UNED. Universidad Nacional de Educación a Distancia, 2002.

[McC01] McCarthy, Diana. *Lexical Acquisiton at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, 2001.

[Oar02] Oard, Douglas W. and Julio Gonzalo. The CLEF 2001 Intercative Track. In Peters, Carol ; Braschler, Martin; Gonzalo, Julio and Kluck, Michael, editor, *Evaluation of Cross-Language Information Retrieval Systems*, pages 309–319. Lecture Notes in Computer Science, Springer, New York, 2002.

[Pet02] Peters, Carol; Martin Braschler, Julio Gonzalo, and Michael Kluck. *Evaluation of Cross-Language Information Retrieval Systems*. Lecture Notes in Computer Science. Springer, New York, 2002.

[Res93] Resnik, Philip S. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, 1993.

[Rol01] Roland, Douglas. *Verb Sense and Verb Subcategorization Probabilities*. PhD thesis, University of Colorado, Colorado, 2001.

[Rol02] Roland, Douglas 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.

[Vos99] Vossen, Piek (ed.). EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document, 1999. http://www.illc.uva.nl/EuroWordNet/docs.html.