Automatic extraction of syntactic semantic patterns for multilingual resources

Borja Navarro, Manuel Palomar, Patricio Martínez-Barco

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, Spain
{borja, mpalomar, patricio}@dlsi.ua.es

Abstract

In this paper we present an automatic system for the extraction of syntactic semantic patterns applied to the development of multilingual processing tools. In order to achieve optimum methods for the automatic treatment of more than one language, we propose the use of syntactic semantic patterns. These patterns are formed by a verbal head and the main arguments, and they are aligned among languages. In this paper we present an automatic system for the extraction and alignment of syntactic semantic patterns from two manually annotated corpora, and evaluate the main linguistic problems that we must deal with in the alignment process.

1. Introduction

One of the most important problems in the development of Natural Language Processing resources is how to deal with the multilingual aspects of communication, that is, how to deal with texts written in different languages. This is a especial problem in Automatic Information Access: in many cases, an information request is encoded in a different language from the one in which the answer is written. The system must process both languages and the semantic relationships between them (López Ostenero, Fernando, 2002).

We think that it is necessary to look for optimum methods for the automatic treatment of more than one language and the relationships between them. In (Navarro et al., 2003b) we presented a proposal for the treatment of multilingual information based on syntactic semantic patterns. According to this proposal, in this paper we present a system that extracts and aligns syntactic semantic patterns from two annotated corpora, one for English (PennTreebank (Marcus et al., 1994)) and the other one for Spanish (the Cast3LB[1]). Both corpus are manually annotated, so we do not deal with problems about the correctness of the syntactic parser. We focus our attention on the pattern extraction process and, specifically, on the bilingual pattern alignment process.

In the next sections we will present our model of syntactic semantic pattern and the acquisition and alignment process. Then the main problems of the alignment process will be discussed. Finally we will present some conclusions and the future works.

2. Multilingual syntactic semantic pattern: definition

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

1. a specific verb (with a specific sense);

2. the syntactic subcategorization frame of the sense: the syntactic category of each verbal argument and the syntactic function of each one;

3. the selectional preferences of each argument: its main semantic features[2].

However, a linguistic pattern in a multilingual information framework must be cross-linguistic. Therefore, together with these three components, a multilingual component is added to the pattern. This component is an identifier of alignment between two or more patterns in different languages. The aligned patterns share the same components (the same verb sense, similar arguments and similar selectional preferences), but they belong to different languages.

The limit of a pattern is the sentence, and its head element is the verb. However, a subcategorization frame is not related directly with a verb (that is, the orthographic word), but it is related directly with the specific sense of the verb ((Roland and Jurafsky, 2002), and others). The verbal subcategorization is based on semantic relationships (Fillmore, 1968), so the arguments are related with the sense of the verb (through a semantic relationship), and not with the orthographic word. This is the main reason why it is necessary to specify the verb sense in the pattern.

On other hand, it is possible to relate a verbal sense with more than one subcategorization frame. It is possible to find two patterns with the same semantic information but with different syntactic information related with the same verb sense. This is the case of diathesis alternation, where the same semantic arguments have two different syntactic configurations with the same verb and in the same context.

From a syntactic point of view, each argument of the pattern is formed by the syntactic category (noun phrase, prepositional phrase, etc.) and the syntactic function (subject, direct object, indirect object, etc.). Due to the system extracts this information from linguistically interpreted corpora, it appears explicitly marked at the corpus with different tags. However, we want to extract only the most closed arguments of the verb. Therefore, the maximum number of arguments extracted are three: subject, object 1 and/or object 2. Other kinds of arguments (locatives, time, etc.) are

---

[1]Project FIT-150-500-2002-244 (Navarro et al., 2003a), (Civit et al., 2003).

[2]For the establishment of this kind of pattern we have take into account several works about subcategorization frame and subcategorization acquisition (Korhonen, 2002), about the relationship between verb sense and verb subcategorization (Roland and Jurafsky, 2002), (Roland, 2001), and about selectional preference acquisition (Resnik, 1993), (McCarthy, 2001).

not considered in this first proposal.

From a semantic point of view, each argument is described by the main semantic features of the phrase head. The semantic features that describe the arguments are extracted from the 63 Top Concepts of EuroWordNet (Vossen, 2002). These Top Concepts have been created from a cross-linguistic point of view, in order to use it for all the languages involved in EuroWordNet.

However, we think that these concepts are too specific for the semantic characterization of arguments. Our objective is to extract general patterns that will be applied to unrestricted domains, so the semantic description of the arguments must be general too. In fact, there are some semantic distinctions among the top concepts of EuroWord-Net without any relation to the argument distinction among monolingual or bilingual patterns. For example, there are not differences between an argument NP with the semantic feature "animal" and other argument NP with the semantic feature "creature". Related to the same verbal sense and without any other argument, they will form the same syntactic semantic pattern.

On other hand, the semantic features must be enough specific to characterize syntactic functions. For example, the semantic distinction [+-animate] is useful in Spanish to distinguish between a NP subject and a NP Object for a verb in active form.

For these reasons, and based on (Chomsky, 1965), we try to use only six semantic features, structured in three couples: *[+abstract/-abstract [-animate/+animate [-human/+human]]]*. Therefore, the 63 higher-level concepts have been reduced to these six general semantic features.

These semantic features are more general than the ones used in other databases of verbal patterns like VerbNet, PropBank or Framenet. In some cases, the semantic features that they apply are quite specifics and in others they are more generals to verb classes (Kingsbury and Palmer, 2002). We prefer the use of general semantic features because we are looking for the maximum recall, and with these general features we avoid the extraction of semantic distinctions related to only one language.

Finally, the bilingual alignment between two patterns from different languages is marked with an alignment index: the identification number of the related pattern. According to the alignment process, the relationship between two aligned patterns is always transitive.

For example, from these two sentences:

(SPN) "Tu hermano se comió los garbanzos de la cocina"
(ENG) "He eats apples every day",

the system extracts patterns like the ones represented on Table 1 and Table 2.

3. The pattern extraction process

As we said before, our main objective is to work in a multilingual framework: to develop an automatic pattern extraction system applied to several languages. However, nowadays, our main task is to detect the automatic alignment problems between patterns extracted from two dif-

Comer:				Id: S1.1
Sbj	hermano	NP	[+ human]	
Obj1	garbanzo	NP	[- animated]	
				AligIndx: E1.1

Table 1: Spanish syntactic semantic pattern

To eat				Id: E1.1
Sbj	he	NP	[+ human]	
Obj1	apple	NP	[- animated]	
				AligIndx: S1.1

Table 2: English syntactic semantic pattern

ferent languages. For this reason, we are working only in a bilingual framework with two corpora: one for English and one for Spanish. The English corpus is the PenTree-bank (Marcus et al., 1994) and the Spanish corpus is the Cast3LB (Navarro et al., 2003a)(Civit et al., 2003). Both corpora are manually annotated with part of speech, syntactic categories and functions tags. This information is directly extracted from the corpora.

However, at this moment, these corpora have not been annotated with semantic information[3]. This is the reason why it is necessary to consult an external resource (EuroWordNet) in order to extract the semantic information.

The extraction process is organized in four steps:

1. The first step is the localization and extraction of the verb.

2. Once the verb is extracted, the system looks for the arguments of the verb. In this step, the system follows a specific order (first the subject, second the object 1 and finally the object 2) looking for the function tag. For each argument located, the nominal, pronominal or verbal head is extracted, together with the function tag and the syntactic category.

   If a specific function tag is not located, the system looks for the next argument, and no argument is extracted for this function.

   On other hand, if a function tag is located but no nominal, pronominal or verbal head is located inside, the system extracts an empty argument, that is, an argument with a specific function but without information about syntactic category or semantic features. This unspecified argument will align with any other argument with the same function.

   Finally, there are some sentences in which it is possible to locate an argument with a verbal head. This is the case of subordinate clauses. In these cases, the head of the argument is the verb. However, this kind of argument is a new pattern too. Therefore, for the same verb the system extracts an argument and a pattern.

---

3. The next step is the extraction of semantic features for each head argument. Once a nominal or verbal head is extracted, the system searches their semantic features in the collection of features derived from EuroWord-Net.

   The main problem in this step is the semantic ambiguity of the words: in many cases, it is necessary to disambiguate the sense of the head. However, due to we have few semantic features (only six), the ambiguity of the words is reduced. There are a lot of words with two or more senses in which all these senses share the same semantic features. Therefore, for our aims, there is not ambiguity. Only in cases where a word has more than one sense with different semantic features for each one, the system extracts the most general feature, because it includes the specific one.

4. The last step is to count the occurrences of each extracted pattern. With this step, the system stores statistical information about the patterns.

   Others systems that extract subcategorization frames use the statistical information in order to refine the extracted frames, and delete all the frames with low occurrences in the corpus (Korhonen, 2002). However, we have not designed a refinement process due to the kind of corpora used. First, the system extracts the arguments according with the function tags marked on the corpus (so the extraction errors are lower than in the use of not annotated corpora). Second, the manually annotated corpora do not have enough occurrences of some verbs. Therefore, if all verbs with low occurrences are deleted, it is possible to delete correct patterns. For example, in the Cast3LB corpus there are 1325 different verbs: 211 verbs have only two occurrences in the corpus, and 587 only one occurrence. Finally, a pattern with low occurrences but with an aligned pattern in other language is a correct pattern. For these reasons, it is possible to delete erroneous patterns after the alignment process only.

4. Pattern alignment process

Once the patterns of each language have been extracted in two individual databases, the alignment process begin. The process is based on the comparison of patterns. Two patterns from different languages align if they share specific information. Therefore, the correctness of the alignment process depends on the correct specification of this information. It is basically semantic information, because it is more abstract and language independent than the syntactic information. However, the syntactic function is used for the alignment process too.

Two patterns from different languages align if they share the next kind of information:

- The same verb sense.

- The same number of arguments.

- The same semantic features of each argument. Together with the sense of the verb, this is the main criterion for the alignment process: our hypothesis is that two patterns from different languages with the same semantic features in each argument are the same pattern at ontological level.

- The same syntactic function of each argument (subject, direct object, etc.). We include this syntactic criterion in the alignment process because we want to align syntactic semantic patterns, not only semantic patterns. That is, we want to align patterns with a specific syntactic configuration.

  As we said before, it is possible to extract a semantic pattern with more than one syntactic configuration. Our objective is to know which of these syntactic configurations are common in both languages. However, if the system uses all the syntactic information of the pattern (syntactic category, function, etc.), the precision and recall of the alignment process will fall. For this reason, we only use the syntactic function in the alignment process, but not the syntactic category. That is more language independent than the syntactic category.

5. Alignment problems

The main linguistic problems that we have found in the alignment process are the next:

- In order to specify the verb sense, a sense disambiguation process is necessary. The use of word sense disambiguation systems generate some errors, that are reflected at the final alignment process.

- The recall of the alignment process is low because the corpora are not very large.

- From sentences with an elliptic argument, the system extracts patterns without an argument. In this cases, the patterns do not align with their related patterns in the other language because they do not share the same number of arguments.

- Some patterns with the same semantic information in each argument do not align because the syntactic function is not shared. There are several cases for this problem.

  A first case is the diathesis alternation. For example, in the case of active-passive alternation, the semantic information of each argument is the same, but the syntactic function is different. If the same semantic pattern appears in active form in one corpus and in passive form in the other one, these patterns do not align.

  A second case is that, sometimes, both languages do not share the same syntactic function for the same pattern. For example, sentences in English like:

  (ENG) "John likes fruit"

  are related with sentence in Spanish like:

  (SPN) "A Juan le gusta la fruta".

  In English, the argument [+human] is the subject, but in Spanish is the object; and in English the argument [-animate] is the object, but in Spanish it is the subject.

The patterns extracted from these sentences are the same, but they do not align because they do not share the same syntactic function in each argument.

- Finally, there are some problems related to the semantic configuration of the languages. For example, there are verbs that in one language have an internal argument, but in the other language have an external argument. In this cases, the same verb has different number of arguments in each langauge. Therefore, the patterns do not align.

  For example , the English verb "to pencil" is usually used in Spanish with the complex form "escribir con un lápiz" ("to write with a pencil"), where the instrumental argument "lápiz" ("pencil") is an external argument. Other example is the verb "to spoon" in English, "echar con una cuchara" ("to throw with a spoon") in Spanish.

6. Conclusions and future works

In this paper we have presented the basic lines of an automatic syntactic semantic pattern extraction system from manually annotated corpora, a cross-linguistic alignment process, and the main linguistic problems that must be solved.

This is a first prototype of a database of cross-linguistic aligned patterns. Nowadays we are solving these alignment problems with the use of other linguistic resources. In future works we will extract patterns from not annotated corpora, and we will apply the method of alignment to these new patterns. Our aim is to extend the database and to solve problems of recall. Other topic that we want to deal with is the specification of the semantic role. With this new information, some problems of alignment will be solved too.

7. Acknowledgements

8. References

Chomsky, Noam, 1965. *Aspects of the Theory of Syntax*. Massachusetts Institute of Technology.

Civit, Montserrat, M[a] Antonia Martí, Borja Navarro, Núria Bufí, Belén Fernández, and Raquel Marcos, 2003. Issues in the Syntactic Annotation of Cast3LB. In *4th International Workshop on Linguistically Interpreted Corpora (LINC03), EACL03*. Budapest.

Fillmore, Charles, 1968. The case for case. In E. Bach and R. Harms (eds.), *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, pages 1–88.

Kingsbury, Paul and Martha Palmer, 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas (Spain).

Korhonen, Anna, 2002. *Subcategorization acquisition*. Technical Report. Cambridge: University of Cambridge.

López Ostenero, Fernando, 2002. *Un sistema interactivo para la búsqueda de información en idiomas desconocidos por el ususario*. Ph.d. diss., Departamento de Lenguajes y Sistemas Informáticos. Universidad Nacional de Educación a Distancia, Madrid.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger, 1994. The penn treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*.

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

Navarro, Borja, Monserrat Civit, María Antonia Martí, Raquel Marcos, and Belén Fernández, 2003a. Syntactic, semantic and pragmatic annotation of Cast3LB. In *Shallow Processing of Large Corpora. A Corpus Linguistics workshop*.

Navarro, Borja, Manuel Palomar, and Patricio Martínez-Barco, 2003b. A general proposal to multilingual information access based on syntactic-semantic patterns. In Antje Düsterhöft and Bernhard Thalheim (ed.), *Natural Langauge Processing and Information Systems - NLDB 2003*. pages 186–1999.

Resnik, Philip S., 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Roland, Douglas, 2001. *Verb Sense and Verb Subcategorization Probabilities*. Ph.D. thesis, University of Colorado, Colorado.

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

Vossen, Piek, 2002. EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document. http://www.illc.uva.nl/EuroWordNet/docs.html.