Syntax to Semantics Transformation: Application to Treebanking

Manuel Alcántara
Universidad Autónoma de Madrid
manuel@maria.lllf.uam.es

Antonio Moreno
Universidad Autónoma de Madrid
sandoval@maria.lllf.uam.es

Abstract

Mapping between syntax and semantics is one of the most promising research topics in corpus annotation. This paper deals with the implementation of an semi-automatic transformation from a syntactically-tagged corpus into a semantic-tagged one. The method has been experimentally applied to a 1600-sentence treebank (the UAM Spanish Treebank). Results of evaluation are provided as well as prospective work in comparing syntax and semantics in written and spoken annotated corpora.

1 Introduction

This paper presents a new stage in the development of the UAM Spanish Treebank[1] (syntactic annotation) and of SESCO[2] (semantic annotation), based on possible relationships between them. Our aim is to achieve semi-automatic semantic tagging of the UAM Spanish Treebank. To reach this goal, we have developed and implemented a program called SST (Syntax Semantics Transformation). The application of this tool provides us with three main benefits:

First and foremost, our principal concern is a reliable, quick and cost effective tagging of the treebank. Manual annotation would be time consuming and expensive because of the complexity of the sentences. On the other hand, automatic processing ensures coherence and control over the tagging: each type will be tagged always with the same structure. Our previous experience in developing the UAM Spanish Treebank has led us to this approach.

Second, this experiment contributes to the study of the relationship between syntax and semantics showing that an almost automatic transition from one level to another is possible. The idea of the connection between these levels of the language is a commonplace in contemporary linguistics and there are important theoretical works concerning the mapping between morpho-syntactic and semantic forms. However these theories have not yet been applied to corpora. Indeed, we are not aware of any similar experiments.

Finally, through this research we have developed a set of grammatical rules connecting syntactic structures to their correspondent event types. It is worth mentioning here that we have worked with corpora with no thematic restrictions. Therefore, these rules are not thought for a particular sublanguage, but applicable to Spanish language in general.

2 UAM Spanish Treebank and SESCO

In order to understand the SST, it is interesting to consider the characteristics of the corpora we have used.

2.1 UAM Spanish Treebank (source corpus).

The UAM Spanish Treebank of the Universidad Autónoma de Madrid is a syntactically annotated corpus made up of 1600 sentences taken from Spanish newspapers (Moreno *et al.*, 1999; Moreno *et al.*, 2003).

Since these sentences (particularly the first 500) were chosen as a sample of the complexity of Spanish syntax, they cover an important range of syntactic structures. The fact that the sample was taken selectively from different sections of the sources reflecting different styles implies much more complexity.

The format was based on the Penn Treebank, although the tag set has been adapted to the characteristics of the Spanish language. The corpus has recently been

The research of Manuel Alcántara has been supported by a grant of the Spanish Ministerio de Educación y Ciencia (FPU).

[1] http://www.lllf.uam.es/~sandoval/UAMTreebank.html and Moreno et al. (2003).

[2] http://www.lllf.uam.es/~manuel/sesco.htm and Alcántara (2003).

converted to an XML format, which has helped ~~us~~ a lot in our work.

The Treebank has four different types of information:

1. Part-of-Speech (noun, verb, etc.)
2. Syntactic functions (SUBJ, DO, ATTR, etc.)
3. Morpho-syntactic features (gender, number, person, etc.)
4. Semantic features. The UAM Spanish Treebank has a group of tags called "semantic features" which specify types of prepositional phrases (locative, time, etc.)

The aim of this annotation was to reflect the surface syntax. The designers were thus very cautious in regards to empty categories and ambiguities: they used the features only in those cases with the highest certainty. Additionally,, the designers avoided redundancy as much as possible.

The Treebank tag set has a flexible design allowing the addition of new features. However as more features are added, annotation becomes more difficult, since the human tagger has to choose the suitable tag among the available ones.

2.2 SESCO (target corpus).

SESCO is a tagging system which allows the semantic representation of a linguistic corpus (Alcántara, 2003). It is coded using an XML markup and offers a practical basis for tagging both spoken and written corpora.

The main goal of SESCO is to make an essential and flexible analysis for extracting the largest possible amount of data from a corpus without limiting it to an excessively restrictive theory, taking the argument structure of verbs as starting-point.

We back J.C. Moreno's proposal (J.C. Moreno 1991a, 1991b, 1997) on event analysis, although we also have considered other very similar approaches (Pustejovsky, 1995; Tenny and Pustejovsky, 2000).

The events expressed by verbs can be of three major types, forming a universal hierarchy (J.C. Moreno, 1997): states, processes and actions. These three types are divided into subtypes according to the arguments they require.

This approach is compositional: a state has two arguments, a process is made up of a transition from one state to another, and an action is a process with an agent. This leads to the logical consequence that we need an annotation format for representing both the relation between events and the arguments of the sentence and its sub-event structure.

Most of the recent work on semantics focuses on ontologies. It is important to distinguish the fact that SESCO does not have an ontology as a basis, but that the ontology can be a result of our work.

SESCO has been developed taking as point of reference the spoken corpus from the Computational Linguistics Laboratory of the Universidad Autónoma de Madrid (http://www.lllf.uam.es/), which, in turn, forms part of the European project "C-ORAL-ROM" (http://lablita.dit.unifi.it/coralrom/). Texts have been recorded following requirements of spontaneity, quality of the sound and variety of speakers and contexts.

At the beginning of our experiment, 49500 spontaneous spoken words (4100 sentences) had been analyzed in SESCO format. These sentences are our training corpus and the basis of our SESCO Data Base (SDB) of event structures.

2.3 Main differences.

Besides the linguistic background, there are three main differences between the syntactically annotated UAM Treebank and SESCO:

First, whereas the Treebank is a corpus of written texts, SESCO contains only spontaneous speech orthografic transcriptions. As we expected, the vocabulary was not the same and the upshot of this was an increase in the number of unknown lemmas. In actual fact, both corpora are designed for covering a wide range of topics and registers.

Second, the UAM Treebank tagset is far more complex than that of SESCO. In this respect, the SST process is a reduction and it does not use all the features included in the Treebank. Syntactic functions and some semantic features are the only information that SST makes use of.

Finally, SST raises fundamental questions on the concept of 'sentence'. In the Treebank, the key is the orthography: the limits of a sentence are always established by dots. In SESCO, a sentence is a complete event. Because of this, the 1600 sentences of the UAM Spanish Treebank corpus produce 1666 sentences in the SESCO version. In spite of this, orthographic punctuation has been helpful in the task of recognizing the beginning of most of the sentences.

	Sentences	Words	Events
UAM Treebank	1666	23542	2230
training SESCO corpus	4100	49506	6530
TOTAL	5766	73048	8760

Table 1 Relevant figures in the corpora

3 Methodology

The input is a syntactically annotated sentence and the output is the same sentence semantically tagged. Both annotations are in XML and the involves five main stages. The first three stages are automatic, implemented in Perl. The fourth (optional) stage is semi-automatic and the last one is ~~a~~ human-revision.

3.1 Getting the event type.

As pointed out earlier, our semantic tagging reflects argument structures related to verbs. Due to this theoretical framework, the first step is to find the lemma of the main verb. It is an easy task since the treebank format provides this information through a particular attribute ("lemma") in the element "verb".

Once the lemma is found, the program searches the SDB for the most frequent event type for this lemma. This selection is made taking into account the syntactic structure: for example, if it is a process and there is a locative complement, the most used displacement will be chosen.

The SDB data come from the previous analysis (for more details about the SESCO corpus, see section 2.2.). That is, this stage is based on a probabilistic model and the automatic mapping is example-based, finding similar examples already in the training corpus.

3.2 From a syntactic structure to a semantic analysis.

In order to understand this second step, first of all it is necessary to remark on some characteristics of the UAM Spanish Treebank. When the UAM Treebank was designed in 1997 (Moreno et al., 2003), the aim was only to build a syntactically annotated corpus following the Penn Treebank style – no consideration was given to the possibility of its translation into a semantic corpus. Therefore the Treebank included only those features needed for achieving a correct syntactic analysis. As mentioned above, the UAM Spanish Treebank uses the standard Penn Treebank scheme with the addition of some features. It provides a combination of Part of Speech information with specific grammatical features of words and phrases.

In SST, this syntactic data is transformed into an event analysis through application of a set of rules. Each rule corresponds to the most frequent correlation between a syntactic phrase and a part of the event structure. Some of the rules are general, but others depend on the lemma. In the current version, lemmas are classified into six different groups:

1. Standard-Type. The rules are consistent with most of the lemmas. By way of illustration, these rules transform the subject (SUBJ) of a sentence, which corresponds to an action, into the agent, and the direct object (DO) into the patient. If the event type is a state, the SUBJ will be the first argument of the state and the attribute will be the second argument. There is a subset of rules for passive sentences.
2. First-Type-Actions. The rules transform the indirect object (IO) into the patient. For instance, “pegar” (*to hit*).
3. Second-Type-Actions. The IO is transformed into the first argument of the states. For instance, “devolver” (*to give back*).
4. Third-Type-Actions. The DO is transformed into the second argument of the states. For instance, “otorgar” (*to grant*).
5. First-Type-States. The IO is transformed with the second argument of the states. For instance, “gustar” (*to like*).
6. Second-Type-States. The second argument of the state is a prepositional phrase. For instance, “coincidir con” (*to coincide with*).

3.3 References and variables.

Lemmas of complex events (specifically actions) are classified additionally depending on their references. References are used in SESCO in order to link the arguments of an event with their functions in the arguments of sub-events. As we have seen in section 2.2, SESCO is based on a compositional semantic theory where actions and processes are made up of sub-events. These references are determined in the case of actions by five different types of lemmas.

Those parts of the event structure which have no correspondence with a phrase (for instance, the agent in a sentence without explicit SUBJ) are filled with variables by the program.

3.4 Unknown lemmas.

As mentioned, the method requires a database with previous examples, something which is not available for all the potential lemmas of a language. In case the program could not reach a model for a lemma, it prompts the user for the most basic information and tries to carry out the analysis. By this means, the final file contains all the sentences in SESCO format with the most likely structure.

Since SESCO has a DTD-controlled tagset covering all possible analysis, the output file will always be a well formed and valid XML file.

3.5 Revision.

The last step is a manual revision of the output file. As we have used the tagging of the UAM Spanish Treebank in order to develop our system, this step has a great importance.

The program errors detected during the analisys have served us to implement new rules. That is why the corpus has been tagged in small groups of sentences (with approx. 100 sentences each group).

When an error is detected during the analysis, typically a new rule is added. For this reason, the corpus has been tagged in small groups of sentences (with approx. 100 sentences each group). Thus, we have

performed sixteen re-examinations of our system each time re-testing the reliability of the rules.

Once the revision is completed, the new sentences are added to the SDB.

4 Main problems for SST.

The last step of the SST process, the revision, provides us with a typology of problems in the automatic part of the system. Let us look at the four most important types and at the number of errors in the 1666 sentences:

Total sentences	Missing lemmas	Verb Type	False Analysis	Treebank errors	*Total*
1666	69	71	66	53	259
100%	4.14%	4.26%	4%	3.18%	15.58%

Table 2. Error typology

1. Sentences without lemmas (69 errors). Newspapers have a lot of sentences (words between dots) which do not have a verb. Nominalization is frequently used by journalists with pragmatic functions. Taking into account that we are analysing argument structures of verbs, this sentence serves to illustrate this *error*: "Medidas desesperadas en China para frenar la crecida del Yangtzé en la provincia de Hubei." ("Tough measures in China to stop the Yangzte overflow in Hubei").
2. Verb Type (71 errors). The analysis of the verb is not correct because it is not in its right group (see section 3.2.). When the SST program does not recognize a lemma, it asks for the essential information, but it does not ask for types of references.
3. False analysis (66 errors). The most likely analysis (according to the SDB) does not correspond to the sentence. Since we are still developing SESCO, it would be naïve to suppose that all these errors are due to SST problems. As we have seen, the SDB is based on a small corpus of 49500 words and they are not enough to get the most likely structure of some verbs (some of which have appeared only once or have not appeared at all).
4. Treebank errors (53 errors). We began our work with the last 100 sentences of the UAM Spanish Treebank (sentences 1500-1600). We have done it in this inverse order because Manuel Alcántara had annotated himself the last sentences of the Treebank. In this process, we have noticed differences between the analysis of the sentences. These differences, even though they are not important for the syntactic analysis, have hindered the SST process since our program expects a particular structure. With the help of SST, we now have a revised version of the syntactic Treebank.

In addition to these errors, there are others which we have not considered so important because they do not change the event type.

The rules for the indirect relations (those phrases which are not arguments of the verb) depend on the semantic features of the Treebank tagset and they are not always enough to determine the right tag. It is worth remembering that both systems (Treebank and SESCO) are designed independently.

Telicity of events is determined by the (indefinite/definite) articles of the phrases. When the head of a phrase is not at the very beginning, errors can occur.

5 Examples.

Let us point out an uncomplicated example of the SST process: "EEUU tiene ya pistas sobre el doble atentado en Kenia y Tanzania ." ("The United States already has a lead about the terrorist outrage of Kenya and Tanzania").

First of all, SST searches for the main verb and its lemma. In this case, the verb is "tiene" (has) and the lemma is "tener" (to have). The Treebank tag for this verb is:

<V Lemma="tener" Tensed="Yes" Form="PRES" Mode="IND" Number="SG" P="3">tiene</V>

From this starting-point, SST looks for the most likely structure of "tener" in the SDB. 99.5% of "tener" events are attributive states with a possessor and a property.

The program checks if "tener" belongs to a special verb type. It does not, so the program checks if it is a normal sentence (it is not in passive voice) and follows the standard rules. These rules are the following:

1. The subject of the sentence ("EEUU") is the possessor.
2. If there is an attributive phrase or a direct object, it is the property. If there is not, the program looks for other possibilities (oblique complement, predicative complement, clauses and prepositional phrases). In our example, "pistas sobre el doble atentado en Kenia y Tanzania" is tagged as direct object.
3. In case no possessor or property was found, SST would assign a variable to these arguments.

4. The program checks if the arguments are definite or indefinite. “Pistas” is indefinite and SST sets the event as indefinite.
5. Finally, SST looks for indirect relations (prepositional phrases which are not arguments).

Once these rules are applied, the program determines if it is a negative sentence, a question, etc. by means of looking for negative words and punctuation, and sets the appropriate features. It also determines the tense.

At the end, the final version of the sentence analysis is written in a target file following the SESCO format.

To take a more difficult example, let us analyze the sentence “Se ha escapado de casa” (“*He/she* has escaped from *his/her* home”). We have only one previous analysis of the lemma “escaparse” (to escape) in SDB and it is an action made up of a displacement.

Regarding references, “Escaparse” belongs in a particular group of events together with “ir”, “irrumpir”, “marchar”, “presentarse”, etc. For this group, the agent and patient of the action and the first argument of the displacement’s states are the same entity.

The SST checks if it is a normal sentence and follows the fitting rules for this group:

1. The subject of the sentence will be the agent. In this case, there is no subject and the program establishes a variable (X) chosen arbitrarily.
2. Because it is a displacement, SST looks for prepositional phrases with “de” or “desde” (“from”) in order to fill the second argument of the first state. It finds “de casa”.
3. SST looks for prepositional phrases with “a” or “hasta” (“to”) in order to fill the second argument of the second state. It does not find it.
4. The program establishes a number as identifier of the agent and links it together with the patient and the first arguments of the states.
5. SST looks for indirect relations.

At last, the program determines that it is not a negative sentence and gets time and mood information.

The annotated sentence in Treebank and SESCO formats can be found in appendix. Most important data is underlined.

6 Future work.

Once we have the UAM Spanish Treebank semantically annotated, we would like to compare the data from both spontaneous speech and written corpora.

In a first comparison, we found that actions are the most frequent event type in our written corpus while states are the most frequent in the spoken one.

		states	processes	actions
Written corpus	total	871	220	1139
	%	39.1%	9.9%	51.1 %
Spoken corpus	total	3939	478	2113
	%	60.3%	7.3%	32.4%

Table 3 Event type comparison

In addition, we are trying to carry out a reverse SST process for achieving a syntactic tagging based on our semantic schemes. On the first stage, we have added morphological information (POS and grammar features as genre, number, etc.) to our SESCO corpora. We want to explore this and other features in future work.

References

Alcántara, Manuel. 2003. "Semantic Tagging System for Corpora". *Proceedings of the Fifth International Workshop on Computational Semantics IWCS-5*. 442-445.

Moreno, Juan Carlos.1991a. *Curso universitario de lingüística general. Tomo I: Teoría de la gramática y sintaxis general.* Síntesis, Madrid.

Moreno, Juan Carlos. 1991b. *Curso universitario de lingüística general. Tomo II: Semántica, Pragmática, Fonología y Morfología.* Síntesis, Madrid.

Moreno, Juan Carlos. 1997. *Introducción a la lingüística. Enfoque tipológico y universalista.* Síntesis, Madrid.

Moreno, Antonio, López, Susana, Sánchez Fernando. 1999. *Spanish Tree bank: Specifications. Version 4. 30 April 1999*. Internal document, Laboratorio de Lingüística Informática, UAM.

Moreno Antonio, López Susana., Sánchez Fernando, Grishman Ralph.. 2003. Developing a syntactic annotation scheme and tools for a Spanish Treebank. In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*. Kluwer, Dordrecht. 149-163.

Pustejovsky, J. 1995. *The Generative Lexicon.* MIT press, Massachussets.

Tenny C. and Pustejovsky, J. 2000. *Events as Grammatical Objects.* CSLI Publications, California.

Appendix.

"SE HA ESCAPADO DE CASA" ("He/she *has escaped from* his/her *home*")

UAM Treebank source sentence:

```
<Sentence Number= "90">
<NP Function= "SUBJ" Number= "SG" P= "3" Elided= "Yes"/>
 <VP Tense= "Tensed" Verbal_temp= "PRES" Aspect= "PERFECT" Mode= "IND" Number= "SG" P= "3" Coordinated= "Yes">
   <VP Tense= "Tensed" Verbal_temp= "PRES" Aspect= "PERFECT" Mode= "IND" Number= "SG" P= "3">
          <NP>
           <P Lemma= "se" Type= "PERS" P= "3" Discontinuous= "Yes" Ref= "1">Se</P></NP>
           <V Verbal_Temp= "ha escapado" Lemma= "escaparse" Tensed= "Yes" Form= "PRES" Mode= "IND" Aspect="PERFECT" Number= "SG" P= "3">
            <AUX Lemma= "haber" Tensed= "Yes" Form= "PRES" Mode= "IND" Number= "SG" P= "3">ha</AUX>
             <V Lemma= "escaparse" Tensed= "No" Form= "PART" Gender="MASC" Number= "SG" Clitic= "Yes" Discontinuous="Yes" ID="1">escapado</V></V>
          <PP Type= "DE" Class= "LOCATIVE">
            <PREP Lemma= "de">de</PREP>
            <NP>
             <N Lemma= "casa" Type= "Common" Gender= "FEM" Number= "SG">casa</N></NP></PP></VP>
<PUNCT Type= "PERIOD" /></Sentence>
```

SESCO target sentence:

```
<S N="90">
   <TEX> 1068-Se 1069-ha 1070-escapado 1071-de 1072-casa</TEX>
         <E TE="action" TYPE="affecting" SUBTYPE="atelic">
                  <LEX LEM="escaparse" VAL="positive" MO="declarative" TI="past"> 1068-Se 1069-ha 1070-
   escapado</LEX>
                  <ARG>
                           <AG IDE="_1">(X)</AG>
                           <PA REF="_1"></PA>
                  </ARG>
                  <E TE="process" TYPE="displacement" SUBTYPE="atelic">
                           <LEX LEM="escaparse" VAL="positive"/>
                           <E TE="state" TYPE="locative" SUBTYPE="indefinite">
                                    <LEX LEM="to be" VAL="positive"/>
                                    <ARG>
                                             <POS REF="_1"></POS>
                                             <LOC IDE="_2">1071-de 1072-casa</LOC>
                                    </ARG>
                           </E>
                           <E TE="state" TYPE="locative" SUBTYPE="indefinite">
                                    <LEX LEM="to be" VAL="negative"/>
                                    <ARG>
                                             <POS REF="_1"></POS>
                                             <LOC REF="_2"></LOC>
                                    </ARG>
                           </E>
                  </E>
         </E>
   </O>
```