Defining a framework for the analysis of predicates

Montserrat CIVIT
University of Barcelona
mcivit@ub.edu

M.Antònia MARTÍ
University of Barcelona
amarti@ub.edu

Roser MORANTE
Tilburg University
r.morante@uvt.nl

Borja NAVARRO
University of Alicante
borja@dlsi.ua.es

Mariona TAULÉ
University of Barcelona
mtaule@ub.edu

Izaskun ALDEZABAL
University of the Basque Country
jibalroi@si.ehu.es

Abstract

In this position paper we present the research on verb predicates that we have carried out until now for Catalan, Spanish, and Basque, and we outline the framework of our future research, which is based on the idea that it is necessary to include syntagmatic and statistic information in lexical resources, such as WordNet, in order to use it in tasks of information extraction from annotated corpora, and in automatic syntactic and semantic tagging of corpora.

1 Introduction

The main goal of this position paper is to summarize the work on verb predicates that we developed in the last years from several perspectives: lexical semantics, corpus linguistics, and the semantic-syntax interface. Starting from that, and taking into consideration recent developments in the field, we sketch a working framework for the automatically semantic tagging of corpora taking advantage of the existing (but limited) semantic resources we have, and of the syntactic information they contain.

In section 2 we explain the motivations of our work, in section 3 we describe the work made until now and we evaluate it, and in section 4 we put forward a framework of future research.

2 Setting

This proposal is the result of several years of research in lexical semantics and corpus linguistics. In our analysis of Spanish, Catalan, and Basque verbs we found that studies about predicates in the fields of lexical semantics and the semantics-syntax interface show a lack of adjustment between the theoretical linguistic analysis and the real problems arising from automatic corpus processing.

Languages do not behave equally regarding the process of automatic analysis. The fact that English is a fixed constituent–order language allows a better adjustment between the theoretical description expressed in the computational lexicon and grammar, and the texts to be analyzed. Catalan, Spanish, and Basque are free constituent–order languages. This characteristic makes more complex the treatment of basic sentences and of the position of arguments. This is why the development of NLP tools, basically lexicons and grammars, has produced relatively poor results and has made evident the mismatch between theoretical work and real samples of language.

As it has been proposed in the literature, lexicons and grammars apart from containing linguistically motivated theoretical information should also incorporate information about coocurrence frequencies and about collocations. Our hypothesis is that adding this information makes the resources more efficient as NLP tools. Thus it is necessary to fill the gap between lexical resources and corpora by enriching lexical resources with syntagmatic information extracted from real samples of language.

3 Background

The authors of this article have worked in individual and in common projects related to NLP from different approaches: lexical semantics, analysis of predicates, and corpus analysis. In what follows a general description of these research directions is presented with the aim of justifying and establishing the basis for future work.

3.1 Lexical semantics

The knowledge sources and the training corpus for Basque, Catalan, and Spanish that were used in the Senseval–2 and Senseval–3 competitions have been developed adopting a lexical semantics perspective.

The quality of lexical resources used in the development of tagged corpus is one of the aspects that has been less taken into account in the Senseval competition. We carried out an experiment in which four different resources were evaluated: Minidir (Márquez et al., 2004), DRAE, EuroWordNet, and a dictionary based on the proposal by (Veronis, 2001). The aim of this experiment was to evaluate the quality of the resources and its effects in the results of the disambiguation tasks. The experiment consisted in letting the same corpus be annotated by three different annotators with each of the three dictionaries. The starting hypothesis was that the higher agreement between annotators would determine which was the lexical resource with more quality. As a result of this research it was found that the corpora with highest degree of agreement between annotators were the corpora annotated with the dictionary elaborated following (Veronis, 2001)'s model[1] and with MiniDir, a dictionary elaborated specifically for the Senseval competition, in which a criterium of minimum granularity of senses was applied. The average of senses per entry is 4. The degree of agreement in these cases was 90%. Both DRAE and EWN gave quite lesser results, between 60% and 70% of agreement.

In Senseval–3 the groups that had worked on Spanish showed a considerable improvement with respect to those of Senseval–2. Later it was found that the cause of the improvement was the methodology applied in the elaboration of the training corpus (Márquez et al., 2004): the same disambiguation system achieves better results if it is trained with the Senseval–3 corpus, than if it is trained with the Senseval–2 corpus.

In the context of Senseval–3, annotators were asked about the linguistic knowledge they were using when assigning senses to words. There was general agreement in considering that the syntagmatic information contained in MiniDir about collocations was essential, as well as the examples. Besides, all of them coincided in considering that the disambiguation of nouns and adjectives was resolved by looking at the strict local context, whereas for verbs it was necessary to identify subject and object.

From all this we deduce, firstly, that the syntagmatic information plays a main role in the disambiguation process, as already pointed out by (Veronis, 2001). Secondly, it seems necessary to have information about the subject and object in order to identify the sense of a verb. In (Nica et al., 2004) it is shown that the definition of syntactic patterns in a corpus and the extraction of paradigmatic information from them brings the corpus closer to the lexical resource and improves the quality of WSD systems.

3.2 Analysis of predicates

With the aim of elaborating a manual verbal classification based on syntactic–semantic criteria, we carried out several studies in the line of (Levin, 1993). The goal was to identify the diathesis in Catalan, Spanish, and Basque (Aldezabal, 2004), and to define verb semantic classes.

This work showed that establishing basic criteria to explain the relation between the diathesis and the verb senses is not straightforward, and that there are difficulties in distinguishing between diatheses and syntagmatic configuration. We produced a list of basic diatheses for 1.200 verbs of Catalan and Spanish, and for 100 verbs in Basque. In the case of Catalan and Spanish the 1200 verbs were grouped in two big classes: verbs of change (which accept the anticausative alternation) and verbs of transfer (which accept the underspecification of the trajectory component).

In the case of Basque, the same theoretical perspective was adopted, but instead of restricting the analysis to some verb classes, each verb was analyzed taking into consideration its occurrences in the corpus, as well as other diatheses alternations that Basque allows. This study showed that for syntactic alternations to have semantic classification power it is necessary to define in a declarative way what is an alternation, and the semantics it reflects (why certain structures form an alternation, which roles or semantic components do participate in it, which are the syntactic phenomena to take into account). Otherwise, when trying to identify the alternations for every verb in the corpus, doubts appear from the very first example.

In sum, those and recent studies for other languages (Schulte im Walde and Erk, 2005) showed that the problem of semantic verb classification, far from being solved, was becoming more complex, and that behind that problem lies the sense disambiguation problem.

3.3 Work with corpus

The authors have collaborated in the development of three treebanks with syntactic and semantic information (3LB corpus). The treebanks are three corpus of 100.000 words for Catalan, Spanish, and Basque, syntactically tag-

[1] Only four entries were elaborated following this model.

ged with phrases and functions. A subset of the corpus has been semantically tagged with WordNet (Palomar et al., 2004). In the Spanish 3LB corpus all nouns, verbs, and adjectives have been tagged. The Cast3LB corpus has approximately 1.400 verbs. In the Catalan corpus the same has been done for 10.000 words.

It is well known how difficult the elaboration of semantically tagged corpus is and how much human effort it costs. Although the data available are sparse, 3LB constitutes the first attempt at providing these languages with a corpus annotated with syntactic and semantic information.

The relation between the senses of a verb and the WordNet senses for subjects and objects can be automatically extracted from the 3LB corpus. Additionally, it is also possible to obtain the syntagmatic structures associated with every verb sense. Data sparseness is the problem that arises in this case, since in order to extract syntactic-semantic information the quantity of examples available is insufficient.

In addition to the 3LB corpus, the Spanish, Catalan, and Basque corpus developed for the lexical sample task in the Senseval–2 and Senseval–3 competitions are available. This corpus has 200 examples for each of the 50 words selected for the lexical sample task, which sums up a total of 10.000 sentences with only 1 tagged word. 10 of the 50 words are verbs (2.000 exemples).

4 Research lines

As it has been said above, we consider that for lexical resources to be useful in language analysis tasks (parsing), as well as in WSD task, it is necessary to enrich them with syntagmatic information. This is what the experiments carried out for tagging the corpus show. The problems that arise are how to acquire this knowledge automatically or semiautomatically, while at the same time guaranteeing its quality, how it will be coded later, and how it will be used.

It is our purpose to develop basic resources for Spanish, Catalan, and Basque in order to provide necessary tagged corpora that will allow carrying out machine learning experiments. In what follows, we propose some strategies based on automatic methods to create some of those resources.

4.1 Semantic disambiguation

Taking as a basis the material that we already have (projects 3LB and Senseval–3), our main goal is to perform an experimental study about the possible correlation between verb senses and semantic type of objects. The information obtained in this way will be added to shallow parsed corpora.

In order to do that we will start from the 3LB corpus (100.000 words), which has both syntactic (functions) and semantic information (synsets of WordNet) for the categories noun, verb, and adjective. For all the senses of verbs with a frequency rate higher than 20, we will extract the noun acting as head of the direct object (DO) and the associated synset. After that we will obtain its *specification mark* (Montoyo, 2002) for the list of direct objects of each sense. This is to say, we will find out in WordNet the lower synset (hypernym) that includes all or most of the synsets associated with the heads of the DO. Our hypothesis is that this node or specification mark defines, for every sense, a subset of EWN where candidates to DO can be found.

In order to verify the relevance of the results obtained, we will check in the Senseval-3 corpus (where only verbs are annotated with synsets) if there exist heads of NPs in the subset of EWN defined by the specification mark. If the result is positive, it will be considered a positive proof in the verification of the hypothesis. If the result is negative, the corpus Senseval–3 will be annotated syntactically, following the methodology defined for 3LB, with the aim of obtaining evidence about the correlation between the verb sense and the semantic type of the object. In the last case a wide collection of examples for every verb is available (a minimum of 200 examples), that might provide more evidence about the validity of the starting hypothesis.

If the results are positive, for the analyzed verbs it will be possible to use the resulting information with the following purposes: assigning the syntactic function *DO* to the NPs of shallow parsed corpora; assigning synsets to all DO on the basis of the specification mark; and semantically tagging the analyzed verbs.

We do not consider making the same study with subjects because in the three languages the subject is usually omitted, and because it is less determining in the semantics of verbs. A next step would be to analyze the prepositional arguments.

4.2 Enriching EWN with syntagmatic information

The information obtained from the previous experiments can be inserted in EWN. For every verbal synset it would be possible to express the nominal synsets that appear as an object, so that this information can be used in WSD processes.

4.3 Syntax–semantics interface

The 3LB corpus provides the necessary information to find out if it exists a correlation between syntactic structures and verbal senses. For example, for each main verb, the Cast3LB and Cat3LB corpora provide information like the specific sense of the verb, the main complements related to this verb, the kind of phrase, the syntactic function of the complements related to the verb, the head of each complement, and its specific sense. So, from these corpora it is possible to extract syntactic semantic patterns formed by each verb and their arguments (Navarro et al., 2004), and it is possible to develop a lexical data base of verb patterns.

Furthermore, we have designed a method for the interlingua alignment of patterns (based on the Interlingua Index of EuroWordNet), in which each pattern is related to the patterns of the same verb sense in other language. This method compares the semantic and syntactic features of each argument of each verb sense, and aligns them if there is syntactic and semantic consistency. With this approach, the diatheses problem is extended to a multilingual framework: indeed, one of the main problems in multilingual alignment of syntactic semantic patterns is that there are different diatheses alternations in different languages (Navarro et al., 2004).

This information can be helpful to complement the work already done about verb diatheses in Basque, Catalan, and Spanish. Starting from this basis it is possible to carry out a translinguistic study about how each language solves the expression of a diathetic expression.

5 Conclusions

In this position paper we have presented a methodology (4.1 and 4.2) for the syntactic and semantic tagging of corpora using information extracted from the 3LB multilingual treebank, and from the automatic analysis of predicates (4.3) of the three languages involved.

6 Acknowledgements

Part of the research presented in this paper has been financed by the following projects: Xtract2 (BFF2002–04226–C03–03) y 3LB (FIT–150500–2003–411).

References

I. Aldezabal. 2004. *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz.* PhD thesis, Basque Philology Department.University of the Basque Country, Leioa.

B. Levin. 1993. *English Verb Classes and Alternations.* Chicago University Press, Chicago.

LL. Márquez, M. Taulé, M.A. Martí, N. Artigas, M. García, F. Real, and D. Ferres. 2004. Senseval–3: The spanish lexical sample task. In *Senseval–3 Third international Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 47–52. ACL, East Stroudsbourg PA–USA.

A. Montoyo. 2002. *Desambiguación léxica mediante marcas de especificidad.* PhD thesis, Univerity of Alacant, Alacant.

B. Navarro, M. Palomar, and P. Martínez-Barco. 2004. Automatic extraction of syntactic semantic patterns for multilingual resources. In *Proceedigns of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon.

I. Nica, M.A. Martí, A. Montoyo, and S. Vázquez. 2004. Combining ewn and sense-untagged corpus for wsd. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 5th International Conference, CICLing 2004 Seoul*, pages 188–200. Springer–Verlag, Heidelberg.

M. Palomar, M. Civit, A. Díaz, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, M.A. Martí, and B. Navarro. 2004. 3lb:construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano. *Procesamiento del Lenguaje Natural*, 33:81–88.

S. Schulte im Walde and K. Erk. 2005. A comparison of german semantic verb classifications. In *Proceedings of the 6th International Workshop on Computational Semantics*, pages 343–353, Tilburg University.

J. Veronis. 2001. Sense tagging: does it make sense? In *Proceedings of the Corpus Linguistics Conference*, Lancaster, UK.