Towards Spanish Verbs' Selectional Preferences Automatic Acquisition. Semantic Annotation of the SenSem Corpus

Jordi Carrera(**), Irene Castellón(*), Salvador Climent(**), Marta Coll-Florit(**)

Grup de Recerca Interuniversitari en Aplicacions Lingüístiques (GRIAL)
(*) Departament de Lingüística General. Universitat de Barcelona
(**) IN3. Internet Interdisciplinary Institute. Universitat Oberta de Catalunya

(*) Facultat de Filologia, Edifici Josep Carner, 5a planta	(**) IN3, Parc Mediterrani de la Tecnologia, Edifici B3
Gran Via de les Corts Catalanes, 585	Avinguda del Canal Olímpic, s/n.
08007 Barcelona	08860 Castelldefels (Barcelona)
Spain	Spain

E-mail: jordi@lightforge.com, icastellon@ub.edu, scliment@uoc.edu, mcollfl@uoc.edu

Abstract

We present the results of an agreement task carried out in the framework of the KNOW Project and consisting in manually annotating an agreement sample totaling 50 sentences extracted from the SenSem corpus. Diambiguation was carried out for all nouns, proper nouns and adjectives in the sample, all of which were assigned EuroWordNet (EWN) synsets. As a result of the task, Spanish WN has been shown to exhibit 1) lack of explanatory clarity (it does not define word meanings, but glosses and examplifies them instead; it does not systematically encode metaphoric meanings, either); 2) structural inadequacy (some words appear as hyponyms of another sense of the same word; sometimes there even coexist in Spanish WN a general sense and a specific one related to the same concept, but with no structural link in between; hyperonymy relationships have been detected that are likely to raise doubts to human annotators; there can even be found cases of auto-hyponymy); 3) cross-linguistic inconsistency (there exist in English EWN concepts whose lexical equivalent is missing in Spanish WN; glosses in one language more often than not contradict or diverge from glosses in another language).

Introduction

The goal of the research we present here is to develop a quality-oriented resource for Natural Language Understanding (NLU): a human-checked verb lexicon of Spanish which includes Selectional Preferences (SPs) based on EuroWordNet's (Vossen, 1998) lexical concepts (*synsets*) and linked, at the same time, to a hand-made semantically-annotated corpus, SenSem (Castellón et al., 2006).

This work is being carried out in the framework of the KNOW project[1], which is aimed at the acquisition of syntactic patterns (Alonso et al., 2007) and SPs from corpora. SPs are a key resource in parsing and Word Sense Disambiguation (WSD): combining syntactic and semantic information allows to refine sentence interpretation, while SPs can be used at the same time for assigning syntactic-semantic dependencies (Atserias, 2006).

In this paper we present the first phase of this project, oriented to assessing annotators' performance with respect to the structure and quality of the lexical-semantic resources used. The paper is divided as follows: first we present related work (Section 1); then we comment on the procedure followed for carrying out the task, as well as on its purpose in the framework of the KNOW project (Section 2); in Section 3, methodology and experimental results of the task are presented, and main sources of disagreement briefly summarized. Section 4, Resource Evaluation, deals with specific cases of disagreement and, based on this discussion, assesses the reliability of wordnets as a tool for NLU. Finally, in Section 5 we present the conclusions of the task and sketch the next steps of our project.

1. Related work

Our work relates mainly to three existing resources: Basque Eusemcor (Agirre et al., 2006), English Propbank (Kingsbury & Palmer., 2002) and Spanish and Catalan AnCora (Martí et al., 2007), while at the same time exhibiting also some characteristic features. Here we will survey these models only briefly. For a more thorough analysis on the topic, readers are addressed to the aforementioned references.

PropBank is the result of adding predicate argument structure annotation to the one-million-words Penn Treebank II Wall Street Journal Corpus (Marcus, 1994). This annotation currently involves verbal predicates only (i.e. nominalizations, adjectives and prepositions have been temporarily excluded). As regards the methodology employed, although it is said that a combination of both semantic and syntactic factors is used, syntactic cues are acknowledged to be foremost: a given predicate's senses are inferred from its usages in the corpus; in fact, for any given usage, "senses" as such are only further specified, later, if required. Syntactic grounds give rise to more corpus-traceable meaning inventories and avoid WordNet fine-grainedness, often associated with arbitrary meaning codification. Finally, it is also worth mentioning that PropBank is intended to be theory-neutral from a semantic standpoint.

Eusemcor is the parallel to English Semcor for Basque language. While carrying out the task of building the Basque WordNet, Agirre et al. (2006), following the advice of Fellbaum et al. (2001), who pointed out that dictionaries focus on word meanings more than on the actual linguistic contexts those meanings are associated with, mirrored the methodology used in the creation of PropBank and undertook the process of simultaneously annotating a Basque corpus while building Basque WordNet, such that the corpus served as a word-usage roadmap to building the knowledge base they had originally set off on creating. This warranted that, upon completion of the work, synsets in Basque WordNet would be consistent with corpus data, on the one hand, while at the same time there being a corpus of word senses manually annotated with respect to Basque WordNet and mapped onto WordNet 1.6.

In order to carry out the twofold process of annotating the corpus while at the same time building the Basque WordNet, the team responsible of developing Eusemcor devised a well-founded methodology (Agirre et al., 2006). In order to carry out the task, an online interface (http://sisx04.si.ehu.es:8080/spsemcor/) was used, one adapted version of which will be used in our research as well. As regards methodology, two remarks must be made: a) taggers tagged independently the same word instances (both in order to measure inter-annotator agreement and also because this might lead to establishing some measure of sense confusability) and b) annotation was carried out transversally instead of linearly (by word-types instead of word-instances -i.e. all corpus instances of a word in one single go-), for this has been shown to enhance agreement measures and annotation consistency (Kilgarriff, 1998).

AnCora (Martí et al., 2007) is a Catalan-Spanish multilingual corpus consisting of two 500,000-words subcorpora. As regards morphology, both corpora have been automatically tagged and manually validated. They have been also syntactically hand-annotated for parts of speech (PoS) and syntactic relations, and the PoS annotation has been used later as a basis for automatically

deriving a dependency-based version of the treebank. As for nouns, both corpora have been manually labeled with WordNet synsets; as regards named entities, both have been manually tagged. For sense-tagging AnCora, guidelines were followed that mostly resemble those adopted by Agirre et al. (2006).

When disambiguating SenSem semantic subcorpus, we will build on all this previous work. At the same time, however, we expect to further theoretical and empirical issues associated with the task. As regards empirical issues, and despite being a monolingual corpus, SenSem will provide a far more balanced coverage (300,000 nouns, verbs and adjectives will have been disambiguated) for Spanish predicates than AnCora, the largest corpus of its kind to our knowledge, which has been semantically tagged only for nouns. Analogously, SenSem is also balanced for verbs (all of them being represented by an average of 100 sentence instances), which will hopefully allow for a systematic study of their predicate argument structure.

From a theoretical point of view, on the other hand, and while also concerned with theory-neutrality, we plan to adopt the approach suggested in (Fellbaum et al., 2001) and already employed in the creation of Eusemcor, by virtue of which corpus occurrences must be the main source of information in order to determine sense distinctions to be mapped onto WordNet, such that synsets express cognitively transparent and statistically significant meanings. Instead of building a new WordNet from scratch, however, we will be primarily concerned with detecting candidates to undergo synset clusterization, for we aim at developing a more compact version of WordNet which is able to reduce noise in sense disambiguation tasks stemming from WN senses' fined-grainedness. In the same vein, while Agirre et al. (2006) weighed the possibility of using meaning distinctions coarser than synsets in order to annotate verbs in WordNet, we intend to broaden sense distinctions and to perform synset clusterization for nouns, with the ultimate goal of arriving at a more evolved and task-efficient version of Spanish EuroWordNet.

To summarize: similar to the approaches of Eusemcor and AnCora, we aim at creating a corpus semantically labeled with synset information. Similar to PropBank, on the other hand, our goal is to create a corpus annotated for predicate structure with which to perform massive argument analysis and focusing on verbs' syntactic frames as a means to study their event structures; differently from their approach, however, we will tag nouns and adjectives and leave verbs aside. Like PropBank and Eusemcor, we will take corpus instances as a departure point for sense-inventory building and lexical knowledge-base modelling. Differently from Eusemcor, however, we will not be building a complete WN anew, for we intend to cluster already existing taxonomic trees in order to come up with more coarse-grained meaning distinctions. As regards methodology, we will import that applied in developing

Eusemcor and AnCora (cfr. inter-annotation, transversal annotation), and we will also be taking advantage of the software developed for building Eusemcor. Finally, mostly like the other approaches, we will adopt a semantically theory-neutral standpoint.

2. Task

The roadmap of the task, along with its short-term developments, can be sketched as follows:

1. Manual annotation of an agreement sample (50 sentences extracted from the SenSem corpus) was performed in order to (a) obtain a human-annotated gold-standard; (b) assess human agreement in the task; and (c) analyze annotators' performance, with the aim of establishing annotation and disambiguation criteria. The annotation consisted in labeling of all nouns, proper nouns and adjectives in the sample, using the appropriate EuroWordNet (EWN) synset. EWN is a multilingual version of WordNet which includes Spanish, Basque, English, Italian and Catalan.

2. The agreement sample was also sense-tagged by an automatic WSD system. Results were evaluated against the hand-made gold-standard, and will ultimately be used as a basis for manual checking and labeling of the SenSem semantic subcorpus (5.000 sentences).

3. Since, as a result of our project, EWN has been also annotated with semantic features (Alvez et al., in press), synset-generalization-based and feature-based SP representations will be compared. Hopefully, this will lead us to a proper theoretical framework for SP representation.

4. Last, we will undertake the process of SP acquisition and, as a result, we will be in a position to implement a Spanish verbs' lexicon.

3. Methodology and experimental results

SenSem is a 700,000-word corpus built up from newspaper articles and containing the 250 most frequent verbs in Spanish, as measured on a wider corpus of newspaper text. SenSem contains an average of 100 examples per verb, all of them annotated syntactically (for phrases and functions) semantically (for eventive structure and theta-roles). Annotated words amount to 300,000.

In the task described here, the agreement sample extracted out of SenSem was PoS tagged using Freeling (Atserias et al., 2006) and semantically tagged using EWN and a WSD system (Cuadros & Rigau, 2007).

Four linguists manually disambiguated the sample. Agreement was deliberately not artificially maximized by enhancing annotators' technical skills before the actual task was carried out (v.gr. taking advantage of annotation guidelines resulting from previous research), but was left to be determined solely on the basis of the empirical data, namely, exclusively considering the words to be labeled

and the resources available to perform the task (i.e. EWN). Error was to be maximized in order to get a better idea of the kind of phenomena that will have to be solved when carrying out a more extensive disambiguation.

The sample was constituted by instances of eight verbs, totaling fifty sentences. Verbs were selected according to variety of a) senses, b) eventive structures and c) semantic classes selected for. For each sentence, nouns, proper nouns and adjectives were manually disambiguated (verbs having been already disambiguated as a result of a previous research under the SenSem Project). Four noun instances were also discarded due to anomalies during extraction. Ultimately, disambiguated words amounted to 185 words over 234 (Table 1). Interestingly, after a first series of disambiguation decisions, inter-annotator agreement was only 40%. The main causes of initial disagreement were the following:

a) Spanish WN's lacking either 1) multiword expressions, 2) metaphoric senses or 3) miscellaneous information.

b) Judges' errors as regards performance during the task or, for some sentences, judges' inability to come up with an interpretation due to lack of contextual information.

Total	Nouns	Verbs	Adjectives
234	67.09%	20.94%	11.97%

Table 1. Word counts for Parts of Speech in SenSem disambiguation sample

The annotation was then discussed in a series of meetings, in order to elicit sources of disagreement and arrive at a consensus on controversial cases such that guidelines for future diambiguation tasks could be properly stablished. This caused agreement to rise to 84,24%. System precision was 42,95% (including monosemic words), as can be seen in the Table 2.

	Agreement	Disagreement
Human	84.32%	15.68%
Machinge	42.95%	57.05%

Table 2. Ratio of human and machine agreement measured on SenSem disambiguation sample

At this point, human disagreement was due to a) noise introduced by too fined-grained meaning distinctions in EuroWordNet (2.5%); b) EuroWordNet's lacking relevant synsets for Spanish (5.6%) and c) disagreement proper between annotators involved (6.7%); only in one case it was the case that a richer context would have been necessary for the target word to be disambiguated. After discussion, some cases remained for which no agreement was reached. For those, the main source of disagreement (10.42%) was EWN's excessive sense granularity

(particularly as regards adjectives). This point is addressed in the following section.

As for the algorithm, we found that there was some margin for improvement as regards the input it received from earlier steps of its functioning: 1% disagreement stemmed from errors in automatic morphological tagging and, likewise, 3.6% disagreement correspondend to undetected multiword expressions whose constituent parts had been wrongly taken separately by the system.

4. Resource evaluation

It is well-known that most WSD systems use wordnets as sense inventories. WordNet senses, however, are characterized by their granularity, so much so that most semantic distinctions are hard to identify even from the point of view of human annotators. This stems from another well-known theoretical problem: what is a word sense? Actually, can words' meaning be split into senses at all? Cognitive linguistics argue that words are radial categories (Taylor, 1995) organized around prototypes, such that it is difficult to distribute them into closed classes (i.e. senses). Kilgarriff (1997) argues that senses can only be inferred from occurrences of words in corpora. These results, however, circumscribe exclusively to the domain for which the clustering is made.

Therefore, WNs' main problem, often pointed out as such, is sense proliferation. It is usually considered that there are too many senses and too alike, which often hinders annotators' task and causes agreement to decrease.

In our sample, recurring cases were found where it was difficult (and maybe even irrelevant for most purposes) to distinguish between:

a) *Senses differing according to perspective or points of view*, e.g. "family" is regarded in EWN either as a group of people who live together, a social group, or a group of people related by marriage and consanguinity. Another case is that of a given word's sense denoting an event, and another sense of the same word referring to the mental perception of that event, e.g. *problem* in *My friends suffer from health problems*: are "health problems" a fact or the observer's conceptualization?

b) *Intrinsically polysemic words*, e.g. an event and the resulting state (as observed for deverbal nouns, e.g. education. When caring about their children's education, do parents care about the way their children are educated, or about the resulting way in which their children behave?).

c) *Senses resulting from meaning modulation due to context*. Only world-knowledge or a richer context would allow one to decide whether (the way it appears in EWN) Spanish "interno" refers to English "boarder", English "inmate" or British English "houseman" (American "intern").

In order to overcome these drawbacks, methods have been proposed that aim at grouping senses (Agirre and Lopez, 2004; Hovy et al., 2006; McCarthy, 2006 and Navigli, 2006). Unfortunately, as yet no freely-distributed resource has been created that covers all of WN, which is why we intend to build a clustered version.

There are, however, some more problems concerning Spanish WN's overall quality that hamper annotation tasks and inter-judge agreement:

4.1. Lack of explanatory clarity

4.1.1. EWN does not use concept definitions, but glosses and examples, for the most part. In many cases, these appear to contradict the meaning the concept should be given attending to taxonomic relationships.

4.1.2. Spanish WN does not systematically collect metaphoric meanings. When this is the case, they must be annotated using the synset corresponding to their literal sense as a compromise solution. Although this is by no means reproachable, since virtually any lexical item can be used in novel ways by virtue of sense extension, it is clearly a conformation measure which postpones the problem rather than solving it, while also triggering new inconsistencies, for it is sometimes the case that WN includes the metaphoric sense of a word without including the literal one, thus rendering the previous solution impractical. As an instance of the first kind of incompleteness, consider "puente" (bridge), which in EuroWordNet appears glossed in the physical structure sense only, whereas in Spanish the noun "puente", through metaphoric extension, can also mean "(air) shuttle service" and "bank holiday". As an instance of the second type of incompleteness, on the other hand, consider "pie" (foot): EuroWordNet includes the "mountain foot" metaphoric reading of "pie", while lacking its literal sense, i.e. "the lower extremity of the vertebrate leg that is in direct contact with the ground in standing or walking".

4.2. Structural inadequacy

4.2.1. Some words appear as hyponyms of another sense of the same word; sometimes there even coexist in Spanish WN a general sense and a specific one related to the same concept, but with no structural link in between; there can even be found cases of auto-hyponymy, which challenge taxonomic principles and are usually symptomatic of deep structural misconceptions: take the case of Spanish "barbacoa" (barbecue). "Barbacoa" is encoded "phonetically", so as to say, that is, hyponymy-hyperonymy links are used to reflect the fact that a single word form can refer either to an artifact (i.e. where meat is roasted), a social gathering (i.e. people meeting for eating food roasted in a barbecue) and food proper (as in "La barbacoa no ha quedado suficientemente hecha", literally "Barbecue is not well-cooked enough"). As it can be clearly seen, all these are variants (i.e. different senses) of the same word, not hyponyms of a single class of

entities which is an event, a roasting appliance and meat at the same time. The first hyponym of this word is actually its literal meaning, all other uses being derived after it. As a matter of fact, we have that a) senses of a word which have been derived metonymyically from it are incorrectly displayed as hyponyms of that word and b) literal meaning is put at a par with metaphoric meaning and encoded as a co-hyponym, such that literal and metaphoric senses' common hypernym must be understood to be a homonymous word form having no particular meaning, for it expresses neither the literal sense (given that this is expressed by one of its hyponyms) nor its related figurative meanings (for these are expressed by the remaining hypoynms), there being no additional meaning left to be coded.

4.2.2. Hyperonymy relationships have been detected that are rather unclear and likely to raise doubts to human annotators. A comprehensive typology of these problems is described in (Guarino 1998). Following his intuitions, we found that "discipline", for instance, was only coded in Spanish WN as an abstract concept. "Discipline", however, while entailing a conceptual meaning component (disciplines certainly do deal with fields of knowledge), is to be better regarded as a human activity (disciplines can lack funding, which is not true as regards concepts, v.gr. theorems). On the other hand, we uncovered some other types of misconceptions Guarino seemed to be unaware of, e.g. hyponymy-meronymy confusion ("bone" appears in Spanish WN as a hyponym of "body tissue", whereas bones are actually MADE-OF body tissue, rather than specific kinds of it).

4.3. Cross-linguistic incoherence

4.3.1. There exist in English EWN concepts that are appropriate in order to hand-annotate our sample corpus, but for which there is no corresponding lexical equivalent in Spanish WN, which forces them to remain unannotated. For example, "juez" (judge) appears in Spanish WN as a monosemic word, only in the legal sense of the term. In Spanish, however, "juez" can also mean "evaluator", which is in fact displayed as an English variant of the Spanish word "juez", but lacks its lexical equivalent in EWN for Spanish.

4.3.2. In multilingual versions of WN, glosses in one language more often than not contradict or diverge from glosses in another language. Consider Spanish "amigo" (friend): whereas the English equivalent is glossed in EWN as "a person you know well and regard with affection and trust", the Spanish term is glossed literally as "a male person you hold friendship with". Likely, the specification "male" has been added to make up for the fact that the Spanish term is inflected for gender. Clearly, however, it has been added wrongly, for in Spanish, as in other languages in which words are inflected for gender, masculine gender is unmarked, that is, it can be used unspecifically to refer to friends of either sex, v.gr. "He salido con unos amigos" means "I've gone out with some friends", rather than "I've gone out with some male friends".

5. Conclusions and future work

We have presented here the work carried out in the task of manually disambiguating an agreement sample extracted from the SenSem corpus. So far, we have been able to complete 1) a training phase, as a result of which annotation and disambiguation criteria have been established, and 2) a second phase where agreement between four human judges, together with performance of an automatic disambiguation system, have been assessed. We have achieved the goal of both detecting problems arising from EWN-based hand-annotation of a sample of the SenSem corpus, as well as that of establishing guidelines for future annotation of the whole database. As future work, our aim is twofold: to develop an appropriate clustering of EWN senses for WSD, on the one hand, and to semantically annotate, on the other, a larger amount of text (SenSem semantic subcorpus), before we run automatic disambiguation.

6. Acknowledgements

References

Agirre E. & Lopez de Lacalle O. (2004). Clustering WordNet Word Senses. In Nicolov et al. (Eds.). Recent Advances in Natural Language Procesing III: Selected Papers from RANLP 2003. John Benjamins Publishers, Amsterdam.

Agirre E., Aldezabal I., Etxeberria J., Iruskieta Quintian M., Izagirre E., Medizabal K. & Pociello E. (2006). Improving the Basque WordNet by corpus annotation. Proceedings of the Third International WordNet Conference, Jeju Island (Korea).

Alonso, L., I. Castellón & N. Tincheva (2007). Obtaining coarse-grained classes of subcategorization patterns for Spanish, Proceedings of the International Conference RANLP.

Álvez J., J. Atserias, J. Carrera, S. Climent, A. Oliver & G. Rigau (in press). Consistent annotation of EuroWordNet with the Top Concept Ontology. In Proceedings of The Fourth Global WordNet Association Conference. Szeged, Hungary.

Atserias, J (2006). Towards Robustness in Natural Language Understanding. Tesi Doctoral. Departamento de Lenguajes y Sistemas Informáticos. Universidad del País Vasco.

Atserias, J., B. Casas, E. Comelles, M. González, L. Padró & M. Padró (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. Proceedings of the Fith International Conference on Language Resources and Evaluation (LREC 2006).

Castellón, I., A. Fernández, G. Vázquez, L. Alonso & J. A. Capilla (2006). The SenSem Corpus: a Corpus Annotated at the Syntactic and Semantic Level, Fifth International Conference on Language Resources and Evaluation (LREC), p. 355-359

Cuadros, M. & G. Rigau (2007). SemEval-2007 Task 16: Evaluation of Wide Coverage Knowledge Resources. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). Prague, Czech Republic.

Fellbaum, C., Palmer, M., Dang H., Delfs L. & Wolff, S. (2001). Manual and Automatic Semantic Annotation with WordNet. In NAACL-2001 Workshop on WordNet and Other Lexical Resources. Pittsburgh, Philadelphia.

Guarino, N. (1998). Some Ontological Principles for Designing Upper Level Lexical Resources. Proceedings of the First International Conference on Language Resources and Evaluation. Granada, Spain.

Hovy, E.H., M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw & R. Weischedel. (2006) OntoNotes: The 90% Solution. Proceedings of the HLT-NAACL 2006, New York.

Kilgarriff, A. (1997) I don't believe in word senses. Computers and the Humanities. Kluwer Academic Publishers, pp. 91-113.

Kilgarriff, A. (1998). Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. In Computer Speech and Language. Special Use on Evaluation 12(4), pp. 453-472.

Kingsbury, P. & M. Palmer (2002). From TreeBank to PropBank. Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas, Spain.

Marcus, M. (1994). The Penn treebank: A revised corpus design for extracting predicate argument structure. Proceedings of the ARPA Human Language Technology Workshop. Princeton, New Jersey.

Martí Antonín, M.A., M. Taulé, M. Bertran, Ll. Màrquez (2007). AnCora: Multilingual and Multilevel Annotated Corpora (in press).

McCarthy, D. (2006). Relating WordNet senses for word sense disambiguation. Proceedings of the ACL Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together . Trento, Italy.

Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING-ACL 2006). Sydney, Australia, July 17-21st, 2006, pp. 105-112.

Taylor, J.R. (1995). Linguistic categorization. Oxford University Press.

Vossen, P. (Ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers