Spanish All-Words Semantic Class Disambiguation Using Cast3LB Corpus[⋆]

Borja Navarro, and Armando Suárez

Departamento de Lenguajes y Sistemas Informáticos.
Universidad de Alicante. Spain
{ruben, loren, borja, armando}@dlsi.ua.es

Abstract. In this paper, an approach to semantic disambiguation based on machine learning and semantic classes for Spanish is presented. A critical issue in a corpus-based approach for Word Sense Disambiguation (WSD) is the lack of wide-coverage resources to automatically learn the linguistic information. In particular, all-words sense annotated corpora such as SemCor do not have enough examples for many senses when used in a machine learning method. Using semantic classes instead of senses allows to collect a larger number of examples for each class while polysemy is reduced, improving the accuracy of semantic disambiguation. Cast3LB, a SemCor-like corpus, manually annotated with Spanish WordNet 1.5 senses, has been used in this paper to perform semantic disambiguation based on several sets of classes: lexicographer files of WordNet, WordNet Domains, and SUMO ontology.

1 Introduction

One of the main problems in a corpus-based approach to Word Sense Disambiguation (WSD) is the lack of wide-coverage resources in order to automatically learn the linguistic information used to disambiguate word senses. This problem is more important when dealing with languages different from English, such as Spanish.

Current approaches to disambiguation using WordNet senses suffer from the low number of available examples for many senses. Developing new hand-tagged corpora to avoid this problem is a hard task that research community tries to solve with semi-supervised methods. An additional difficulty is that more than one sense is often correct for a specific word in a specific context. In this cases, it is hard (or even impossible) to choose just one sense per word.

Using semantic classes instead of WordNet senses provides solutions to both problems [3] [15] [9] [11] [16]. A WSD system learns one classifier per word using the available examples in the training corpus whereas semantic class classifiers can use examples of several words because a semantic class groups a set of senses,

[⋆] This paper has been supported by the Spanish Government under projects CESS-ECE (HUM2004-21127-E) and R2D2 (TIC2003-07158-C04-01).

which are related from a semantic point of view. Therefore, semantic classes allow more examples per class, reduce the polysemy, and allow less ambiguity. Different collections of semantic classes have been proposed. Three of them are used in this paper: lexicographer files (LexNames) of WordNet [5], WordNet Domains (WND) [4] and SUMO ontology [8].

The main goal of this paper is to perform semantic class disambiguation in Spanish, similarly to [15] where several experiments were done with semantic classes for English using SemCor. We used the Cast3LB corpus, a manually annotated corpus in Spanish, for training and testing our system.

The rest of this paper is organized as follows: we first present some previous work related with semantic classes. Section 3 describes the three set of classes used and the Cast3LB corpus. In the next section, section 4, experiments and features are explained. Section 5 shows the results obtained and, finally, some conclusions and futur work are discussed in section 6.

2 Related Work

The semantic disambiguation based on coarse classes rather than synsets is not a new idea. In [9] a method to obtain sets of conceptual classes and its application to WSD is presented. This method is based on the selectional preferences of the verb: several verbs specify the semantic class of its arguments. For example, the selectional preferences of the direct object of a verb like "to drink" is "something liquid".

Other paper that tries to develop semantic disambiguation based on semantic classes is [16]. He uses the Roget's Thesaurus categories as semantic classes.

In [11] LexNames are used in order to automatically learn semantic classes. Its approach is based on Hidden Markov Model.

[15] focuses on the general idea of getting more examples for each class based on a coarse granularity of WordNet. They use LexNames and SUMO ontology in order to translate SemCor senses to semantic classes. They obtain the best results with a reduced features set of the target word: only lemma, PoS and the most frequent semantic class calculated over the training folders of the corpus are used. By using these features they obtain an accuracy of 82.5% with LexNames, and an accuracy of 71.9% with SUMO. According to their results, they conclude that it is very difficult to make generalization between the senses of a semantic class in the form of features.

The aim of [3] is to overcome the problem of knowledge acquisition bottleneck in WSD. They propose a training process based on coarse semantic classes. Specifically, they use LexNames. Once a coarse disambiguation is obtained, they apply some heuristics in order to obtain the specific sense of the ambiguous word (for example, the most frequent sense of the word in its semantic class). They use some semantic features. However, due to the difficulty of making generalization in each semantic class, they do not apply the features as a concatenated set of information. Instead of this, they apply a voting system with the features.

3 Semantic Classes and Cast3LB

In this section, the three sets of classes and the Cast3LB corpus used are described briefly.

3.1 Sets of Semantic Classes

WordNet synsets are organized in forty five lexicographer files, or **LexNames**, based on syntactic categories (nouns, verbs, adjectives and adverbs) and logical groupings, such as person, phenomenon, feeling, location, etc. WordNet 1.5 has been used in the experiments since our corpus is annotated using this WordNet version.

SUMO (Suggested Upper Merge Ontology) provides definitions for general-purpose terms and gathers several specific domains ontologies (such as communication, countries and regions, economy and finance, among others). It is limited to concepts that are general enough to address (at a high level) a broad range of domain areas. SUMO has been mapped to all of WordNet lexicon. Its current version is mapped to WordNet 1.6. and it contents 20,000 terms and 60,000 axioms. It has 687 different classes.

WordNet Domains are organized into families, such as sport, medicine, anatomy, etc. Each family is a group of semantically close SFCs (subject field codes) among which there is no inclusion relation. SFCs are sets of relevant words for a specific domain. Currently, there are 164 different SFCs, organized in a four level hierarchy, that have been used to annotate WordNet 1.6 with the corresponding domains (including some verbs and adjectives).

3.2 The Cast3LB Corpus

In Cast3LB all nouns, verbs and adjectives have been manually annotated with their proper sense of Spanish WordNet, following an all-words approach[1] [6]. Cast3LB examples have been extracted from the Lexesp corpus[10] and the Hermes Corpus[2]. The corpus is made up of samples of different kinds of texts: news, essays, sport news, science papers, editorials, magazine texts and narrative literary texts. In Table 1 statistical data about the corpus is shown. Cast3LB has approximately 8,598 annotated words (36,411 out of 82,795 occurrences): 4,705 nouns, 1,498 verbs, and 2,395 adjectives.

Table 1. Amount of words in the Cast3LB corpus

	Nouns	**Verbs**	**Adjectives**
Occurrences	17506	11696	7209
Words	4705	1498	2395

[1] In an all-words approach, all words with semantic meaning are labelled.

[2] nlp.uned.es/hermes/

Comparing Cast3LB to other corpora annotated with senses, it has a medium size (although Cast3LB is being currently extended up to 300,000 words under the project CESS-ECE). It is smaller than SemCor (250,000 words) [5] and MultiSemCor (92,820 annotated words for Italian) [1]. However, it has more annotated words than the all-words corpora used at Senseval-3 for Italian (5,189 words: 2,583 nouns, 1,858 verbs, 748 adjectives) [13] or English (5000 words approximately) [12].

The corpus has 4,972 ambiguous words out of 8,598, which means that 57.82% of them has two or more senses in Spanish WordNet. The corpus polysemy degree according to each set of classes (WN senses, LexNames, WND and SUMO) is shown in table 2.

Table 2. Polysemy in the Cast3LB corpus

	Senses	**LexNames**	**WND**	**SUMO**
Adjectives	5.69	1.14	2.32	1.27
Nouns	3.84	2.42	2.28	2.95
Verbs	6.66	3.17	2.04	4.35
All	4.91	2.65	2.23	2.96

More information about the annotation process of Cast3LB can be found in[7].

4 Experiments

The experiments have been designed in order to analyze the behaviour of a WSD method based on semantic classes when different sets of classes are used. 10-fold cross-validation has been used and the accuracy for each experiment is averaged over the results of the 10 folds.

4.1 Features

Using information contained in Cast3LB, we want to know how different information affect the disambiguation task based on semantic classes. In this section, we present different kinds of information that have been used in the experiments.

Word Information
This information refers to word form, lemma and PoS. PoS feature allows a coarse semantic disambiguation. Since many words have senses as nouns, verbs or adjectives at the same time, previous knowledge about their PoS tags in some context helps to discard some of such senses. Moreover, Spanish language has a richer morphology than English. So, PoS tags include morphological information as gender and number. In the experiments, we have used this kind of information from target word and words surrounding it.

Bigrams

Words and lemmas within the target word context have been selected to build up the bigrams. Target word is not included in bigrams. With this information we want to find patterns or word co-occurrences that reveal some evidence about the proper class of the target word.

Syntactic Information

Each verb has been marked with their arguments and its syntactic function (subject, object, indirect object, etc.), that is, the subcategorization frame of the verb. So, for each ambiguous word, its syntactic constituents and the syntactic function of the phrase in which it occurs are known, allowing us to use all this information to enrich the set of features.

Topic Information

As topic information, the kind of text in which the target word occurs is used. Cast3LB texts are organized in several folders according to the kind of text: news, sports, etc. The name of such folders is used as an additional feature for the examples extracted from their texts.

In addition, we have used the name of the file as a feature because, in general, different occurrences of a word in the same text tend to have the same sense [18]. This topic information refers only to the target word.

4.2 Description of the Experiments

As said before, we have studied how a WSD system based on semantic classes behaves when different sets of classes are used. Support Vector Machines (SVM) [14] have been selected because their good performance when dealing with high dimensional input space and irrelevant features, as proven in Senseval-3.

The experiments consist of using different kinds of information for each set of classes. The purpose, besides of comparing the performance of the three set of classes, is to reveal which types of features supply relevant information to the learning process by means of excluding them in a particular experiment[3]. Therefore, the experiments are divided into two sets: one set considering only one kind of information for each experiment, and the other set using more than one kind of information. The list of experiments with one type of information is:

- **Word Information (*W*):** word, lema and PoS at -3,-2,-1,0,+1,+2,+3
- **Bigrams (*B*):** word and lemma bigrams at (-3,-2),(-2,-1),(-1,+1),(+1,+2) and (+2,+3)
- **Syntactic Information (*S*):** syntactic function and phrase type of the ambiguous word
- **Topic Information (*T*):** topic information of the target word

[3] We have used a context of 3 words to the left and right of the target word, although Gale, Church and Yarowsky showed that a bigger window is better for class classification. The reason to do so is that we are not interested in reaching the best results, but comparing different semantic classes and kinds of information.

And the list of experiments combining different types of information is:

- **All information (*WBST*):** all the available information is used to train the classifiers. That is: *Word inf.+Bigrams inf.+Syntactic inf.+Topic inf.*
- **Excluding bigrams inf. (*WST*):** it is the same experiment as the *All information* experiment excluding bigrams information. That is: *Word inf.+ Syntactic inf.+Topic inf.*
- **Excluding syntactic inf. (*WBT*)):** not taking into account syntactic information. That is: *Word inf.+Bigrams inf.+Topic inf.*
- **Excluding topic inf. (*WBS*):** we do not use topic information of the target word in this case. That is: *Word inf.+Bigrams inf.+Syntactic inf.*
- **Context (WB_{cont}):** word information at -3,-2,-1,+1,+2,+3 and bigram information of surrounding words.

Notice that no automatic tagging of Cast3LB has been performed but all this information is already available in the corpus. Our main goal is to test the advantages of using semantic classes instead of senses.

5 Evaluation and Results

In this section, the results for each category set (LexNames, WND and SUMO) are shown. The results are separated by PoS and by kind of experiment. Although all semantic annotated words (nouns, verbs and adjectives) have been used to create the classifiers, only nouns and verbs have been selected to test them: in the LexNames set, there are only three possible classes for adjectives; adjective polisemy in SUMO is 1.27, that is nearly monosemic.

As explained before, a SVM learning algorithm has been used, specifically an implementation due to Thorsten Joachims: *SVMLight*[4]. The configuration for the SVM module is simple for a first test: a linear kernel with a 0.01 value for the c regularization parameter.

To verify the significance of the results, one-tailed paired t-test with a confidence value of $t_{9,0.975} = 2.262$ has been used, selecting the results of the $WBST$ experiment as baseline to compare with other experiments. Significant experiments, according to t-test, are highlighted in next tables.

The upper part of Table 3 shows the ordered accuracy[5] values for one kind of information experiments using the three sets of classes and considering only nouns. The ranking for the experiments in the three cases is the same and the bests results are reached by the W experiment using the features: words, lemmas and PoS. The reason is that, while the target word is not usually used for WSD, it plays an important role in semantic class disambiguation. This is so because examples of different words are used to train a semantic class classifier. T and

[4] `svmlight.joachims.org`

[5] All experiments have resulted in 100% of coverage((correct+wrong)/total). In this case, precision(correct/(correct+wrong)) and recall(correct/total) are the same, and are referred as accuracy in this paper.

S experiments have the worst results, because such information is excessively general for certain contexts.

Additionally, results for WND are slightly different than for LexNames and SUMO, mainly because LexNames is a very small set of classes, and the mapping of SUMO and WordNet is done between concepts and concrete senses. WND is more like a sense clustering where each cluster groups a semantically related senses but not necessarily hyperonyms or hyponyms. This results into a different distribution of examples depending on the set of classes.

Table 3. Accuracy for nouns

experiment	LEX	experiment	WND	experiment	SUMO
W	84.09	W	**79.62**	W	81.43
B	**67.72**	B	**69.83**	B	**61.84**
T	**61.86**	T	**67.72**	T	**53.49**
S	**60.47**	S	**63.29**	S	**52.07**
WST	84.7	WST	**83.3**	WST	81.9
WBT	84.4	WBS	82.6	WBT	81.7
$WBST$	84.3	$WBST$	82.5	$WBST$	81.5
WBS	83.8	WBT	**79.8**	WBS	**80.6**
WB_{cont}	**69.4**	WB_{cont}	**71.5**	WB_{cont}	**64.3**

In the bottom part of the same Table 3 results for the experiments with several kinds of information are shown. In order to find out to which extent one kind of information influences the disambiguation results, we compare the results obtained by experiments excluding one kind of information to those obtained by the experiment $WBST$ (using all available information). SUMO and LexNames seem to have the same behaviour while WND is different.

As expected, the worst results are obtained by the **Context** experiments, since they do not contain information about the target word, which is very important in semantic class disambiguation, as we mentioned before.

Syntactic information does not seem to have much influence on semantic disambiguation when using SUMO or LexNames. However, this information seems to play a rol in semantic disambiguation using WND.

Topic information is apparently more relevant for the disambiguation process, for SUMO and LexNames at least. Topic information is useful when combined with other kind of features. Likely, topic information needs to be based on a more sophisticated source than few categories in which the texts are classified. Moreover, text classification tools or even a topic search based on broad context windows will probably provide a more accurate set of features.

Results for verbs are shown in Table 4. As in the previous experiments for nouns, LexNames and SUMO behave in a similar way while WND does not. As we expected, the results for verbs are worse than for nouns, due to the greater polysemy of verbs. An exception is WND where results for verbs are similar to

Table 4. Accuracy for verbs

experiment	LEX	experiment	WND	experiment	SUMO
W	**76.12**	W	87.13	W	68.57
B	**53.14**	B	**86.38**	B	**45.19**
T	**47.67**	T	**86.12**	T	**40.75**
S	**46.23**	S	**85.29**	S	**38.72**
WST	**76.1**	WBT	87.2	WST	**69.0**
WBT	75.4	WST	87.0	WBT	68.7
$WBST$	74.9	WBS	87.0	$WBST$	68.1
WBS	74.6	$WBST$	86.9	WBS	**67.3**
WB_{cont}	**55.7**	WB_{cont}	**86.6**	WB_{cont}	**47.4**

results for nouns because the polysemy for nouns (2.28) and verbs (2.04) in this class set is similar.

Finally, table 5 shows overall results for the disambiguation process taking into account both, nouns and verbs. As expected, the results reflect the same behaviour than considering verbs and nouns separately. Nouns have a bigger impact on SUMO and LexNames, while verbs do on WND. The reason is that verb polysemy is bigger than noun polysemy for SUMO and LexNames. However, noun polysemy is bigger than verb polysemy in the case of WND.

Table 5. Accuracy for nouns and verbs

experiment	LEX	experiment	WND	experiment	SUMO
W	**81.61**	W	**81.96**	W	**77.43**
B	**63.17**	B	**74.99**	B	**56.66**
T	**57.44**	T	**73.45**	T	**49.53**
S	**56.03**	S	**70.14**	S	**47.91**
WST	**82.0**	WST	**84.5**	WST	**77.9**
WBT	81.6	WBS	83.9	WBT	**77.7**
$WBST$	81.5	$WBST$	83.9	$WBST$	77.4
WBS	**81.0**	WBT	**82.1**	WBS	**76.5**
WB_{cont}	**65.2**	WB_{cont}	**76.2**	WB_{cont}	**59.0**

6 Conclusions and Future Work

In this paper, an approach to WSD for Spanish based on semantic classes has been presented. Spanish, as other languages has not many resources for training WSD systems. We have used the Cast3LB corpus, a manually annotated Spanish corpus with WordNet senses.

Some experiments have been carried out in order to study the performance of semantic class disambiguation using three sets of classes: LexNames, SUMO and WordNet Domains. The results are quite similar for each one. Only the results obtained for WND are different.

As the experiments show, the most important information for semantic class disambiguation has to do with the target word. As we have said, examples of different words are used to train a semantic class classifier, and that is why the specific word is so important. On the contrary, others kinds of information and context information are not useful for semantic class disambiguation. Therefore, a more appropriate feature definition must be done for semantic class.

The experiments show that LexNames and SUMO have similar results, while WND behaves in a different way. As stated before, the reason is that LexNames and SUMO are based on WordNet hierarchy. SUMO has been mapped to WordNet 1.6. However, we can conclude that this mapping does not provide any improvement compared to LexNames, since the results for both are quite similar. WND seems to be a more proper resource for semantic class disambiguation in open-domain texts.

At present we are focused on a deeper study of the influence of topic information in WSD based on semantic classes. Although we think that topic information could be useful for semantic class disambiguation, the number of topics we have used does not seem to be large enough. Moreover, we think than topic information can be more useful for the disambiguation of some words than for others. We want to develop a technique to identify those words that are specially affected by topic information. In order to do so, we are testing some threshold techniques to increase precision, labelling only those contexts which are high confidently classified.

Additionally, we are now working on a richer feature definition as well as applying semantic class classification on WSD, Information Retrieval and Question Answering. We are also studying the feasibility of this approach to extract new annotated examples from the Web in order to enlarge Cast3LB.

References

1. L. Bentivogli and E. Pianta. 2005 Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus. *Natural Language Engineering. Special Issue on Parallel Text.*11(3). Pp. 247-261.
2. Montserrat Civit, MA Martí, Borja Navarro, Núria Bufí, Belén Fernández and Raquel Marcos. 2003. Issues in the Syntactic Annotation of Cast3LB. *4th International on Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003 workshop.* Budapest, Hungary.
3. Upali S. Kohomban and Wee Sun Lee. 2005. Learning Semantic Classes for Word Sense Disambiguation. *Proceeding of the 43th Annual Meeting of the Association for Computational Linguistics*, Michigan, USA.
4. Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation.* Athens, Greece.
5. G. A. Miller, C. Leacock, T. Randee and R. Bunker. 1993. A Semantic Concordance *Proceedings of the 3rd ARPA Workshop on Human Language Technology* San Francisco.
6. Borja Navarro, Montserrat Civit, MA Antonia Martí, Raquel Marcos, Belén Fernández. 2003 Syntactic, Semantic and Pragmatic Annotation in Cast3LB. *Corpus Linguistics 2003 Workshop on Shallow Procesing of Large Corpora.*, Lancaster, UK.

7. Borja Navarro, Raquel Marcos and Patricia Abad. 2005 Semantic Annotation and Inter-Annotators Agreement in Cast3LB Corpus. *Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)* Barcelona, Spain.
8. Ian Niles and Adam Pease. 2001 Towards a Standard Upper Ontology *Proceedings of 2nd International Conference on Formal Ontology in Information Systems (FOIS'01)*", Ogunquit, USA
9. Philip Resnik. 1997. Selectional preference and sense disambiguation. *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington D.C., USA.
10. N. Sebastián, M.A. Martí, M. F. Carreiras and F. Cuetos 2000. *LEXESP: Léxico Informatizado del Español* Edicions de la Universitat de Barcelona Barcelona
11. Frederique Segond, Anne Schiller, Gregory Grefenstette and Jean-Pierre Chanod. 1997. An Experiment in Semantic Tagging using Hidden Markov Model Tagging. *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, *Proceedings of ACL 97*, pp. 78-81, Madrid, Spain.
12. Benjamin Snyder and Martha Palmer. 2004 The English All-Word Task. *Porceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* Barcelona, Spain
13. Marisa Uliveri, Elisabetta Guazzini, Francesca Bertagna and Nicoletta Calzolari. 2004 Senseval-3: The Italian All-words Task, *Proceeding of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Anlysis of Texts*, Barcelona, Spain
14. Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory.* Springer.
15. Luis Villarejo, Lluis Márquez and German Rigau. 2005. Exploring the construction of semantic class classifiers for WSD. *Revista de Procesamiento del Lenguaje Natural*, 35:195-202.
16. David Yarowsky. 1992. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.. *Proceedings, COLING-92*, pp. 454-460, Nantes, France.
17. Piek Vossen. 1998. EuroWordNet: a multilingual database with lexical semantic networks for European Languages.
18. W. Gale, K. Church and D. Yarowsky. 1992. One Sense per Discourse.. *Proceedings of the 4th. DARPA Speech and Natural Language Workshop*, pp. 233-237.