SemEval-2007 Task 09: Multilevel Semantic Annotation of Catalan and Spanish

Lluís Màrquez and Luis Villarejo
TALP Research Center
Technical University of Catalonia
{lluism,luisv}@lsi.upc.edu

M. A. Martí and Mariona Taulé
Centre de Llenguatge i Computació, CLiC
Universitat de Barcelona
{amarti,mtaule}@ub.edu

Abstract

In this paper we describe SemEval-2007 task number 9 (*Multilevel Semantic Annotation of Catalan and Spanish*). In this task, we aim at evaluating and comparing automatic systems for the annotation of several semantic linguistic levels for Catalan and Spanish. Three semantic levels are considered: noun sense disambiguation, named entity recognition, and semantic role labeling.

1 Introduction

The Multilevel Semantic Annotation of Catalan and Spanish task is split into the following three subtasks:

Noun Sense Disambiguation (NSD): Disambiguation of all frequent nouns ("all words" style).

Named Entity Recognition (NER): The annotation of (possibly embedding) named entities with basic entity types.

Semantic Role Labeling (SRL): Including also two subtasks, i.e., the annotation of verbal predicates with semantic roles (SR), and verb tagging with semantic–class labels (SC).

All semantic annotation tasks are performed on exactly the same corpora for each language. We presented all the annotation levels together as a complex global task, since we were interested in approaches which address these problems jointly, possibly taking into account cross-dependencies among them. However, we were also accepting systems approaching the annotation in a pipeline style, or addressing any of the particular subtasks in any of the languages.

In Section 2 we describe the methodology followed to develop the linguistic corpora for the task. Sections 3 and 4 summarize the task setting and the participant systems, respectively. Finally, Section 5 presents a comparative analysis of the results. For any additional information on corpora, resources, formats, tagsets, annotation manuals, etc. we refer the reader to the official website of the task[1].

2 Linguistic corpora

The corpora used in this SemEval task are a subset of CESS-ECE, a multilingual Treebank, composed of a Spanish (CESS-ESP) and a Catalan (CESS-CAT) corpus of 500K words each (Martí et al., 2007b). These corpora were enriched with different kinds of semantic information: argument structure, thematic roles, semantic class, named entities, and WordNet synsets for the 150 most frequent nouns. The annotation process was carried out in a semiautomatic way, with a posterior manual revision of all automatic processes.

A sequential approach was adopted for the annotation of the corpus, beginning with the basic levels of analysis, i.e., POS tagging and chunking (automatically performed) and followed by the more complex levels: syntactic constituents and functions (manually tagged) and semantic annotation (manual and semiautomatic processes with manual completion and posterior revision). Furthermore, some experiments concerning inter-annotator agreement

[1] www.lsi.upc.edu/∼nlp/semeval/msacs.html

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 42–47,
Prague, June 2007.

were carried out at the syntactic (Civit et al., 2003) and semantic levels (Màrquez et al., 2004) in order to evaluate the quality of the results.

2.1 Syntactic Annotation

The syntactic annotation consists of the labeling of constituents, including elliptical subjects, and syntactic functions. The surface order was maintained and only those constituents directly attached to any kind of 'Sentence' root node were considered ('S', 'S.NF', 'S.F', 'S*'). The syntactic functions are: subject (SUJ), direct object (OD), indirect object (OI), attribute (ATR), predicative (CPRED), agent complement (CAG), and adjunct (CC). Other functions such as textual element (ET), sentence adjunct (AO), negation (NEG), vocative (VOC) and verb modifiers (MOD) were tagged, but did not receive any thematic role.

2.2 Lexical Semantic Information: WordNet

We selected the 150 most frequent nouns in the whole corpus and annotated their occurrences with WordNet synsets. No other word categories were treated (verbs, adjectives and adverbs). We used a steady version of Catalan and Spanish WordNets, linked to WordNet 1.6. Each noun either matched a WordNet synset or a special label indicating a specific circumstance (for instance, the tag C2S indicates that the word does not appear in the dictionary). All this process was carried out manually.

2.3 Named Entities

The corpora were annotated with both *strong* and *weak* Named Entities. Strong NEs correspond to single lexical tokens (e.g., "[U.S.]$_{LOC}$"), while weak NEs include, by definition, some strong entities (e.g., "The [president of [US]$_{LOC}$]$_{PER}$"). (Arévalo et al., 2004). Thus, NEs may embed. Six basic semantic categories were distinguished: Person, Organization, Location, Date, Numerical expression, and Others (Borrega et al., 2007).

Two golden rules underlie the definition of NEs in Spanish and Catalan. On the one hand, only a noun phrase can be a NE. On the other hand, its referent must be unique and unambiguous. Finally, another hard rule (although not 100% reliable) is that only a definite singular noun phrase might be a NE.

2.4 Thematic Role Labeling / Semantic Class

Basic syntactic functions were tagged with both arguments and thematic roles, taking into account the semantic class related to the verbal predicate (Taulé et al., 2006b). We characterized predicates by means of a limited number of Semantic Classes based on Event Structure Patterns, according to four basic event classes: *states*, *activities*, *accomplishments*, and *achievements*. These general classes were split into 17 subclasses, depending on thematic roles and diathesis alternations.

Similar to PropBank, the set of arguments selected by the verb are incrementally numbered expressing the degree of proximity of an argument in relation to the verb (Arg0, Arg1, Arg2, Arg3, Arg4). In our proposal, each argument includes the thematic role in its label (e.g., Arg1-PAT). Thus, we have two different levels of semantic description: the argument position and the specific thematic role. This information was previously stored in a verbal lexicon for each language. In these lexicons, a semantic class was established for each verbal sense, and the mapping between their syntactic functions with the corresponding argument structure and thematic roles was declared. These classes resulted from the analysis of 1,555 verbs from the Spanish corpus and 1,077 from the Catalan. The annotation process was performed in two steps: firstly, we annotated automatically the unambiguous correspondences between syntactic functions and thematic roles (Martí et al., 2007a); secondly, we manually checked the outcome of the previous process and completed the rest of thematic role assignments.

2.5 Subset for SemEval-2007

The corpora extracted from CESS-ECE to conform SemEval-2007 datasets are: (a) SemEval-CESS-ESP (Spanish), made of 101,136 words (3,611 sentences), with 29% of the corpus coming from the Spanish EFE News Agency and 71% coming from Lexesp, a Spanish balanced corpus; (b) SemEval-CESS-CAT (Catalan), consisting of 108,207 words (3,202 sentences), with 71% of the corpus consistinf of Catalan news from EFE News Agency and 29% coming from the Catalan News Agency (ACN).

These corpora were split into training and test subsets following a a 90%–10% proportion. Each

test set was also partitioned into two subsets: 'in-domain' and 'out-of-domain' test corpora. The first is intended to be homogeneous with respect to the training corpus and the second was extracted from a part of the CESS-ECE corpus annotated later and not involved in the development of the resources (e.g., verbal dictionaries).[2]

3 Task setting

Data formats are similar to those of CoNLL-2004/2005 shared tasks on SRL (column style presentation of levels of annotation), in order to be able to share evaluation tools and already developed scripts for format conversion.

In Figure 1 you can find an example of a fully annotated sentence in the column-based format. There is one line for each token, and a blank line after the last token of each sentence. The columns, separated by blank spaces, represent different annotations of the sentence with a tagging along words. For structured annotations (parse trees, named entities, and arguments), we use the Start-End format. Columns 1–6 correspond to the input information; columns 7 and above contain the information to be predicted. We can group annotations in five main categories:

BASIC_INPUT_INFO (columns 1–3). The basic input information, including: (a) **WORD** (column 1) words of the sentence; (b) **TN** (column 2) target nouns of the sentence, marked with '*' (those that are to be assigned WordNet synsets); (c) **TV** (column 3) target verbs of the sentence, marked with '*' (those that are to be annotated with semantic roles).

EXTRA_INPUT_INFO (columns 4–6). The extra input information, including: (a) **LEMMA** (column 4) lemmas of the words; (b) **POS** (column 5) part-of-speech tags; (c) **SYNTAX** (column 6) Full syntactic tree.

NE (column 7). Named Entities.

NS (column 8). WordNet sense of target nouns.

SR (columns 9 and above). Information on semantic roles, including: (a) **SC** (column 9). Semantic class of the verb; (b) **PROPS** (columns 10 and above). For each target verb, a column representing the argument structure. Core numbered arguments include the thematic role labels. ArgM's are the adjuncts. Columns are ordered according to the textual order of the predicates.

All these annotations in column format are extracted automatically from the syntactic-semantic trees from the CESS-ECE corpora, which were distributed with the datasets. Participants were also provided with the whole Catalan and Spanish WordNets (v1.6), the verbal lexicons used in the role labeling annotation, the annotation guidelines as well as the annotated corpora.

4 Participant systems

About a dozen teams expressed their interest in the task. From those, only 5 registered and downloaded datasets, and finally, only two teams met the deadline and submitted results. ILK2 (Tilburg University) presented a system addressing Semantic Role Labeling, and UPC* (Technical University of Catalonia) presented a system addressing all subtasks independently[3]. The ILK2 SRL system is based on memory-based classification of syntactic constituents using a rich feature set. UPC* used several machine learning algorithms for addressing the different subtasks (AdaBoost, SVM, Perceptron). For SRL, the system implements a re-ranking strategy using global features. The candidates are generated using a state–of–the–art SRL base system.

Although the task targeted at systems addressing all subtasks jointly none of the participants did it.[4] We believe that the high complexity of the whole task together with the short period of time available were the main reasons for this failure. From this point of view, the conclusions are somehow disappointing. However, we think that we have contributed with a very valuable resource for the future research and, although not complete, the current systems provide also valuable insights about the task and are very good baselines for the systems to come.

5 Evaluation

In the following subsections we present an analysis of the results obtained by participant systems in the

[2]For historical reasons we referred to these splits as '3LB' and 'CESS-ECE', respectively. Participants in the task are actually using these names, but we opted for using a more simple notation in this paper (see Section 5).

[3]Some members of this team are also task organizers. This is why we mark the team name with an asterisk.

[4]The UPC* team tried some inter-task features to improve SRL but initial results were not successful.

```
INPUT-------------------------------------------------------------->  OUTPUT-----------------------------------
BASIC_INPUT_INFO----->  EXTRA_INPUT_INFO--------------------------->  NE NS------> SR---------------------->
WORD          TN TV LEMMA       POS                  SYNTAX          NE NS         SC        PROPS---------->
-------------------------------------------------------------------------------------------------------------

Las           -  -  el          da0fp0   (S(sn-SUJ(espec.fp*)         *   -          -           *  (Arg1-TEM*
conclusiones  *  -  conclusion  ncfp000       (grup.nom.fp*           *   05059980n  -           *           *
de            -  -  de          sps00            (sp(prep*)         *   -          -           *           *
la            -  -  el          da0fs0       (sn(espec.fs*)     (ORG*   -          -           *           *
comision      *  -  comision    ncfs000      (grup.nom.fs*           *   06172564n  -           *           *
Zapatero      -  -  Zapatero    np00000         (grup.nom*)     (PER*)  -          -           *           *
,             -  -  ,           Fc                 (S.F.R*          *   -          -           *           *
que           -  -  que         pr0cn00      (relatiu-SUJ*)          *   -          -   (Arg0-CAU*)          *
ampliara      -  *  ampliar     vmif3s0              (gv*)           *   -          a1         (V*)          *
el            -  -  el          da0ms0     (sn-CD(espec.ms*)         *   -          -   (Arg1-PAT*          *
plazo         *  -  plazo       ncms000      (grup.nom.ms*           *   10935385n  -           *           *
de            -  -  de          sps00            (sp(prep*)         *   -          -           *           *
trabajo       *  -  trabajo     ncms000    (sn(grup.nom.ms*)))))     *   00377835n  -           *)          *
,             -  -  ,           Fc                    *))))))    *)  -          -           *           *)
quedan        -  *  quedar      vmip3p0              (gv*)           *   -          b3          *         (V*)
para          -  -  para        sps00          (sp-CC(prep*)         *   -          -           *  (ArgM-TMP*
despues_del   -  -  despues_del spcms             (sp(prep*)         *   -          -           *           *
verano        *  -  verano      ncms000    (sn(grup.nom.ms*))))      *   10946199n  -           *           *)
.             -  -  .           Fp                    *)           *   -          -           *           *
```

Figure 1: An example of an annotated sentence.

three subtasks. Results on the test set are presented along 2 dimensions: (a) *language* ('ca'=Catalan; 'es'=Spanish); (b) *corpus source* ('in'=in–domain corpus; 'out'=out–of–domain corpus). We will use a *language.source* pair to denote a particular test set. Finally, '*' will denote the addition of the two subcorpora, either in the language or source dimensions.

5.1 NSD

Results on the NSD subtask are presented in Table 1. BSL stands for a baseline system consisting of assigning to each word occurrence the most frequent sense in the training set. For new nouns the first sense in the corresponding WordNet is selected. The UPC* team trained a SVM classifier for each word in a pre-selected subset and applied the baseline in the rest of cases. The selected words are frequent words (more than 15 occurrences in the training corpus) showing a not too skewed distribution of senses in the training set (the most predominant sense covers less than 90% of the cases). No other teams presented results for this task.

	All words		Selected words	
Test	BSL	UPC*	BSL	UPC*
ca.*	85.49%	86.47%	70.06%	72.75%
es.*	84.22%	85.10%	61.80%	65.17%
*.in	84.84%	86.49%	67.30%	72.24%
*.out	85.02%	85.33%	67.07%	67.87%
.	84.94%	85.87%	67.19%	70.12%

Table 1: Overall accuracy on the NSD subtask

The left part of the table ("all words") contains results on the complete test sets, while the right part ("selected words") contains the results restricted to the set of words with trained SVM classifiers. This set covers 31.0% of the word occurrences in the training set and 28.2% in the complete test set.

The main observation is that training/test corpora contain few sense variations. Sense distributions are very skewed and, thus, the simple baseline shows a very high accuracy (almost 85%). The UPC* system only improves BSL accuracy by one point. This can be partly explained by the small size of the word-based training corpora. Also, this improvement is diminished because UPC* only treated a subset of words. However, looking at the right–hand side of the table, the improvement over the baseline is still modest ($\sim$3 points) when focusing only on the treated words. As a final observation, no significant differences are observed across languages and corpora sources.

5.2 NER

Results on the NER subtask are presented in Table 2. This time, BSL stands for a baseline system consisting of collecting a gazetteer with the strong NEs appearing in the training set and assigning the longest matches of these NEs in the test set. Weak entities are simply ignored by BSL. UPC* presented a system which treats strong and weak NEs in a pipeline of two processors. Classifiers trained with multiclass

AdaBoost are used to predict the strong and weak NEs. See authors' paper for details.

Test	BSL			UPC*		
	Prec.	Recall	F_1	Prec.	Recall	F_1
ca.*	75.85	15.45	25.68	80.94	77.96	79.42
es.*	71.88	12.07	20.66	70.65	65.69	68.08
*.in	83.06	17.43	28.82	78.21	74.04	76.09
*.out	68.63	12.20	20.72	76.21	72.51	74.31
.	74.45	14.11	23.72	76.93	73.08	74.96

Table 2: Overall results on the NER subtask

UPC* system largely overcomes the baseline, mainly due to the low recall of the latter. By languages, results on Catalan are significantly better than those on Spanish. We think this is attributable mainly to corpora variations across languages. By corpus source, "in-domain" results are slightly better, but the difference is small (1.78 points). Overall, the results for the NER task are in the mid seventies, a remarkable result given the small training set and the complexity of predicting embedded NEs.

Detailed results on concrete entity types are presented in Table 3 (sorted by decreasing F_1). As expected, DAT and NUM are the easiest entities to recognize since they can be easily detected by simple patterns and POS tags. On the contrary, entity types requiring more semantic information present fairly lower results. ORG PER and LOC are in the seventies, while OTH is by far the most difficult class, showing a very low recall. This is not surprising since OTH agglutinates a wide variety of entity cases which are difficult to characterize as a whole.

	Prec.	Recall	F_1
DAT	97.38%	96.88%	97.13
NUM	98.05%	89.68%	93.68
ORG	75.72%	75.36%	75.54
PER	70.48%	75.97%	73.13
LOC	73.41%	68.29%	70.76
OTH	56.90%	37.79%	45.41

Table 3: Detailed results on the NER subtask: UPC* team; Test corpus *.*

Another interesting analysis is to study the differences between strong and weak entities (see Table 4) . Contrary to our first expectations, results on weak entities are much better (up to 11 F_1 points higher). Weak NEs are simpler for two reasons: (a) there exist simple patters to characterize them, without the need of fully recognizing their internal strong NEs; (b) there is some redundancy in the corpus when tagging many equivalent weak NEs in embedded noun phrases. It is worth noting that the low results for strong NEs come from classification rather than recognition (recognition is almost 100% given the "proper noun" PoS tag), thus the recall for weak entities is not diminished by the errors in strong entity classification.

	Prec.	Recall	F_1
Strong NEs	73.04%	63.36%	67.85
Weak NEs	78.96%	78.91%	78.93

Table 4: Results on strong vs. weak named entities: UPC* team; Test corpus *.*

5.3 SRL

SRL is the most complex and interesting problem in the task. We had two participants ILK2 and UPC*, which participated in both subproblems, i.e., labeling arguments of verbal predicates with thematic roles (SR), and assigning semantic class labels to target verbs (SC). Detailed results of the two systems are presented in Tables 5 and 6.

Test	UPC*			ILK2		
	Prec.	Recall	F_1	Prec.	Recall	F_1
ca.*	84.49	77.97	81.10	84.72	82.12	83.40
es.*	83.88	78.49	81.10	84.30	83.98	84.14
*.in	84.17	82.90	83.53	84.71	84.12	84.41
*.out	84.19	72.77	78.06	84.26	81.84	83.03
.	84.18	78.24	81.10	84.50	83.07	83.78

Table 5: Overall results on the SRL subtask: semantic role labeling (SR)

The ILK2 system outperforms UPC* in both SR and SC. For SR, both systems use a traditional architecture of labeling syntactic tree nodes with thematic roles using supervised classifiers. We would attribute the overall F_1 difference (2.68 points) to a better feature engineering by ILK2, rather than to differences in the Machine Learning techniques used. Overall results in the eighties are remarkably high given the training set size and the granularity of the thematic roles (though we have to take into account that systems work with gold parse trees). Again, the results are comparable across languages and slightly better in the "in-domain" test set.

Test	UPC*			ILK2		
	Prec.	Recall	F_1	Prec.	Recall	F_1
ca.*	86.57	86.57	86.57	90.25	88.50	89.37
es.*	81.05	81.05	81.05	84.30	83.63	83.83
*.in	81.17	81.17	81.17	84.68	83.11	83.89
*.out	86.72	86.72	86.72	90.04	89.08	89.56
.	83.86	83.86	83.86	87.12	85.81	86.46

Table 6: Overall results on the SRL subtask: semantic class tagging (SC)

In the SC subproblem, the differences are similar (2.60 points). In this case, ILK2 trained specialized classifiers for the task, while UPC* used heuristics based on the SR outcomes. As a reference, the baseline consisting of tagging each verb with its most frequent semantic class achieves F_1 values of 64.01, 63.97, 41.00, and 57.42 on ca.in, ca.out, es.in, es.out, respectively. Now, the results are significantly better in Catalan, and, surprisingly, the 'out' test corpora makes F_1 to raise. The latter is an anomalous situation provoked by the 'es.in' tset.[5]

Table 7 shows the global SR results by numbered arguments and adjuncts Interestingly, tagging adjuncts is far more difficult than tagging core arguments (this result was also observed for English in previous works). Moreover, the global difference between ILK2 and UPC* systems is explained by their ability to tag adjuncts (70.22 vs. 58.37). In the core arguments both systems are tied. Also in the same table we can see the overall results on a simplified SR setting, in which the thematic roles are eliminated from the SR labels keeping only the argument number (like other evaluations on PropBank). The results are only $\sim$2 points higher in this setting.

Test	UPC*			ILK2		
	Prec.	Recall	F_1	Prec.	Recall	F_1
Arg	90.41	87.73	89.05	89.42	88.58	88.99
Adj	64.72	53.16	58.37	72.54	68.04	70.22
A-TR	92.91	90.15	91.51	91.31	90.45	90.88

Table 7: Global results on numbered arguments (Arg), adjuncts (Adj), and numbered arguments without thematic role tag (A-TR). Test corpus *.*

Finally, Table 8 compares overall SR results on known vs. new predicates. As expected, the results on the verbs not appearing in the training set are lower, but the performance decrease is not dramatic (3–6 F_1 points) indicating that generalization to new predicates is fairly good.

Test	UPC*			ILK2		
	Prec.	Recall	F_1	Prec.	Recall	F_1
Known	84.39	78.43	81.30	84.88	83.46	84.16
New	81.31	75.56	78.33	79.34	77.81	78.57

Table 8: Global results on semantic role labeling for known versus new predicates. Test corpus *.*

Acknowledgements The organizers would like to thank the following people for their hard work on the corpora used in the task: Juan Aparicio, Manu Bertran, Oriol Borrega, Núria Bufí, Joan Castellví, Maria Jesús Díaz, Marina Lloberes, Difda Monterde, Aina Peris, Lourdes Puiggrós, Marta Recasens, Santi Reig, and Bàrbara Soriano. This research has been partially funded by the Spanish government: Lang2World (TIN2006-15265-C06-06) and CESS-ECE (HUM-2004-21127-E) projects.

References

Arévalo, M., M. Civit and M. A. Martí. 2004. MICE: a Module for Named-Entities Recognition and Classification. *International Journal of Corpus Linguistics*, 9(1). John Benjamins, Amsterdam.

Borrega, O., M. Taulé, M. A. Martí. 2007. What do we mean when we speak about Named Entities? In *Proceedings of Corpus Linguistics* (forthcoming). Birmingham, UK.

Civit, M., A. Ageno, B. Navarro, N. Bufí and M. A. Martí. 2003. Qualitative and Quantitative Analysis of Annotatotrs: Agreement in the Development of Cast3LB. In *Proceedings of 2nd Workshop on Treebanks and Linguistics Theories (TLT-2003)*, 33–45. Vaxjo, Sweden.

Màrquez, L., M. Taulé, L. Padró, L. Villarejo and M. A. Martí. 2004. On the Quality of Lexical Resources for Word Sense Disambiguation. In *Proceedings of the 4th EsTAL Conference, Advances in natural Language Processing*, LNCS, vol. 3230, 209–221. Alicante, Spain.

Martí, M. A., M. Taulé, L. Màrquez, and M. Bertran. 2007a. Anotación semiautomática con papeles temáticos de los corpus CESS-ECE. In *Revista de la SEPLN - Monografía TIMM* (forthcoming).

Martí, M. A., M. Taulé, L. Màrquez, and M. Bertran. 2007b. *CESS-ECE: A multilingual and Multilevel Annotated Corpus.* E-pub., http://www.lsi.upc.edu/∼mbertran/cess-ece

Taulé, M., J. Castellví and M. A. Martí. 2006. Semantic Classes in CESS-LEX: Semantic Annotation of CESS-ECE. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT-2006).* Prague, Czech Republic.

[5] By chance, the genre of this part of corpus is mainly literary. We are currently studying how this is affecting performance results on all subtasks and, particularly, semantic class tagging.