Memory-Based Semantic Role Labeling of Catalan and Spanish

Roser Morante and Antal van den Bosch
Dept. of Language and Information Sciences
Tilburg University, P.O.Box 90153
NL-5000 LE Tilburg, The Netherlands
{*R.Morante,Antal.vdnBosch*}@*uvt.nl*

Abstract

In this paper we present a memory-based semantic role labeling (SRL) system for Catalan and Spanish. We approach the SRL task as two distinct classification problems: the assignment of semantic roles to arguments of verbs, and the assignment of a semantic class to verbs. We hypothesize that the two tasks can be solved in a uniform way, for both languages. Building on the same pool of features reported useful in earlier work, we train two classifiers for the two subtasks, selecting features systematically in a hill-climbing search. We use the IB1 classifier, a supervised memory-based learning algorithm based on the *k*-nn algorithm. The system achieves an overall score of 85.69 F-score on Catalan, and 84.12 on Spanish.

Keywords

Semantic role labeling, memory-based learning, Spanish, Catalan.

1 Introduction

Semantic role labeling (SRL) is a sentence-level natural-language processing (NLP) task in which semantic roles are assigned to all arguments of a predicate [9]. Identifying semantic roles can be useful for several NLP applications such as information extraction [15], or in machine translation, where automatically identified predicates can be reordered as a preprocessing step to statistical MT [11]. The CoNLL-2004 and CoNLL-2005 Shared Tasks [1, 2] addressed SRL for English, providing a well-defined context for research and evaluation in this field.

In this paper we present a semantic role labeling system based on a system [14] submitted to the task *Multilevel Semantic Annotation of Catalan and Spanish* [13] in the context of SemEval-2007, providing a detailed analysis of the results obtained. The general SRL task consists of two tasks: the assignment of semantic roles (SR) to arguments of verbs, and the prediction of the lexico-semantic class of the verb (SC). We develop systems for each of the tasks, for each of the two languages. For the SR task there are 39 classes in the Catalan training corpus and 48 in the Spanish training corpus. For the SC task there are 17 classes in both the Catalan and the Spanish corpora. The fact that verbs belong to a certain class depends on their argument structure. For example, class *d2* covers agentive ditransitive verbs, which have a double object (patient, beneficiary), like change of possession (*dar*, 'give') and communication verbs (*decir*, 'tell').

The engine of the two systems for semantic role (SR) and semantic class (SC) prediction for both languages is a memory-based classifier. Memory-based language processing [5] is based on the idea that NLP problems can be solved by storing annotated examples of the problem in their literal form in memory, and applying similarity-based reasoning on these examples in order to solve new ones. Keeping literal forms in memory has been argued to provide a key advantage over abstracting methods in NLP that ignore exceptions and sub-regularities [6]. In general, NLP tasks aiming at aspects of semantic analysis are difficult to model by abstract rules. In SRL, it is difficult to formulate processing rules even for humans because semantic roles are inherently tied to meaning, inheriting all the ambiguity that lexical semantics is faced with – predicates with more than one possible meaning typically license different sets of semantic frames with each meaning. Since lexical word sense disambiguation is shown to be solvable at state-of-the-art levels by memory-based learning [10, 8], and since memory-based learning has also been applied to English SRL [16], we considered using memory-based learning for our present SRL experiments.

Building on a pool of features that have been successfully used in earlier work on SRL, we train two similar classifiers to predict the semantic class of the verb and the semantic roles separately, for both languages. With this study we intend to test whether individual systems could produce competitive results in both tasks, and whether they would be robust enough when applied to two languages and to the out-of-domain test sets provided. Additionally, our goal is to analyse what the most informative sets of features are in this task.

The data provided in the shared task are sentences with tokenized words annotated with lemmas, parts-of-speech, syntactic information, semantic roles, and the semantic classes of the verb. Although the setting is similar to the CoNLL-2005 Shared Task, two important differences are that the corpora are significantly smaller (500K words), and that the syntactic information is based on a manually annotated treebank carrying information on syntactic functions (i.e. direct object, indirect object, etc.).

```
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                      *)          *  -           -           *           *
```

Fig. 1: *An example of an annotated sentence [13].*

For additional information on the corpora, tagsets, and annotation manuals, we refer the reader to [13], and to the official website of the task[1].

The paper is organised as follows. In Section 2 we present our double-classifier SRL system. In Section 3 the results are presented at various levels of granularity, and we report on feature selection experiments. In Section 4 we formulate our conclusions.

	Training 3LB		Test 3LB		Test CESS	
	Ca.	Sp.	Ca.	Sp.	Ca.	Sp.
SR	23202	24668	1335	1451	1241	1186
SC	8932	9707	510	615	463	465

Table 1: *Number of instances per corpus for each task ('Ca' stands for Catalan, 'Sp' stands for Spanish).*

2 System description

We approach the SRL task as two distinct classification problems: the assignment of semantic roles to arguments of a predicate (SR) and the assignment of semantic classes to predicates (SC). We hypothesize that the two problems can be solved uniformly for both languages. We build two similar systems that differ only in some of the features used, as outlined below.

Both the SR and the SC tasks are solved in two phases: (1) A pre-processing phase of *focus selection*, similar to the sequentialization step in [12]. Focus selection consists of identifying the potential candidates to be assigned a semantic role or a semantic verb class. (2) Classification.

Regarding the focus selection process, the system starts by detecting a target verb, which is marked in the corpora as such. Then it identifies the complete form of the verb (which in the corpus is tagged as verb group, infinitive, gerund, etc.) and the clause boundaries in order to look for the siblings of the verb that exist within the same clause. The phrases with syntactic function subject are annotated in the corpora as siblings of the verb. For each sentence, the focus selection process produces two groups of focus tokens: on the one hand, the verbs, and on the other, the siblings of the verbs. These tokens will be the focal elements of the examples in each training set. Table 1 shows the number of training and test instances for each task.

We approach the SR and SC tasks as single-step classification tasks. We assume that all verbs belong to a class, so we generate one classification for each verb. As for the SR task, we assume that most siblings of the verb will have a class, except for those that have syntactic functions AO, ET, MOD, NEG, IMPERS, PASS, and VOC, as these never carry a semantic role in the training corpora. The siblings that do not have a semantic role are assigned the NONE tag. Because the corpus is small and because the amount of instances with a NONE class is proportionally low, we do not consider it necessary to filter these cases out.

Regarding the learning algorithm, we use the IB1 classifier as implemented in TiMBL (version 5.1) [7], a supervised inductive algorithm for learning classification tasks based on the k-nearest neighbor classification rule [4]. In IB1, similarity is defined by a feature-level distance metric between a test instance and a memorized example. The metric combines a per-feature value distance metric with global feature weights that account for relative differences in importance of the features.

In our study the IB1 algorithm was parametrized by using Jeffrey Divergence as the similarity metric, the modified value-difference metric [3] as the per-feature value distance metric, gain ratio for feature weighting, using 11 k-nearest neighbors, and weighting the class vote of neighbors as a function of their inverse linear distance [7].

We developed the systems by performing cross-validation experiments, iterated for every step in the feature selection process. Feature selection was performed by starting with a set of basic features (essentially the identity and the parts-of-speech tags of the head words involved, in their local context) and gradually adding new features. For training and testing,

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

the systems took a total average of 9 seconds for the SC task (i.e. 64.38 instances per second), and 87 seconds for the SR task (16.04 instances per second). To evaluate the systems, two test sets were used for each language: one from the same training corpus (3LB) and one out-of-domain test set (CESS–ECE).

2.1 Features

We collected a pool of features that should in theory be useful for both the SC and SR tasks. Most of these features are described in earlier work as providing useful information for semantic role labeling [9, 18, 1, 2, 17]. These features encode the identity and other syntactic aspects of the verb in focus and its clausal siblings. After experimenting with 323 features, we selected 98 for the SR task and 77 for the SC task. In order to select the features, we started with a basic system, the results of which were used as a baseline. Every new feature that was added to the basic system was evaluated in terms of average accuracy in a 10-fold cross-validation experiment; if it improved the performance on held-out data, it was added to the selection. One problem with this hill-climbing method is that the selection of features is determined by the order in which the features have been introduced. We selected it because it is a fast heuristics method, in comparison, for example, with genetic algorithms. We also performed experiments applying the feature selection process reported in [16], a bi-directional hill climbing process. However, experiments with this advanced method did not produce a better selection of features.

The features for the SR prediction task are the following (we put in brackets the number of features):

- Features of the verb in focus (6). They are shared by all the instances that represent phrases belonging to the same clause:

  VForm; **VLemma**; **VCau**: binary features that indicate if the verb is in a causative construction with *hacer, fer* or if the main verb is *causar*; **VPron**, **VImp**, **VPass**: binary features that indicate if the verb is pronominal, impersonal, and in passive form respectively.

- Features of the sibling in focus (12):

  SibSynCat: syntactic category; **SibSynFunc**: syntactic function; **SibPrep**: preposition; **SibLemW1, SibPOSW1, SibLemW2, SibPOSW2, SibLemW3, SibPOSW3**: lemma and POS of the first, second and third words of the sibling; **SibRelPos**: position of the sibling in relation to the verb (PRE or POST); **Sib+1RelPos**: position of the sibling next to the current phrase in relation to the verb (PRE or POST); **SibAbsPos**: absolute position of the sibling in the clause.

- Features that describe properties of the content word (CW) of the focus sibling (10): in the case of prepositional phrases, the CW is taken to be the head of the first noun phrase; in cases of coordination, we only select the first element of the coordination.

  CWord; **CWLemma**; **CWPOS**: we take only the first character of the POS provided; **CWPOSType**: the type of POS, second character of the POS provided; **CWGender**; **CWne**: boolean feature that indicates if the CW is a named entity; **CWtmp, CWloc**: boolean features that indicate if the CW is a temporal or a locative adverb respectively; **CW+2POS, CW+3POS**: POS of the second and third words after CW.

- Features of the clause containing the verb in focus (24):

  CCtot: total number of siblings with function CC; **SUJRelPos, CAGRelPos, CDRelPos, CIRelPos, ATRRelPos, CPREDRelPos, CREGRelPos**: relative positions of siblings with functions SUJ, CAG, CD, CI,ATR, CPRED, and CREG in relation to verb (PRE or POST); **SEsib**: boolean feature that indicates if the clause contains a verbal *se*; **SIBtot**: total number of verb siblings in the clause; **SynFuncSib8, SynCatSib8, PrepSib8,W1Sib8, W2Sib8, W3Sib8, W4Sib8, SynFuncSib9, SynCatSib9, PrepSib9, W1Sib9, W2Sib9, W3Sib9, W4Sib9**: syntactic function, syntactic category, preposition, and first to fourth word of siblings 8 and 9.

- Features extracted from the verbal frames lexicon (43). The task organization provided lexicons of verbal frames for Catalan and Spanish. We access the lexicon to check if it is possible for a verb to have a certain semantic role:

  The features are boolean: **Arg0–AGT**, **Arg0–CAU**, **Arg0–EXP**, **Arg0–TEM**, **Arg1–AGT**, **Arg1–PAT**,**Arg1–TEM**,**Arg2–ATR**,**Arg2–PAT**,**Arg3–ATR**, **ArgM–CAU**, **Arg2–LOC**, **Arg2–ADV**, **Arg1–LOC**, **Arg3–LOC**,**ArgM–ADV**,**ArgM–LOC**,**ArgM–MNR**, **ArgM–TMP**, **Arg0**, **Arg1**, **Arg1–EXT**, **Arg2**, **Arg2–BEN**, **Arg2–EFI**, **Arg2–EXT**, **Arg2–INS**, **Arg2–ORI**, **Arg3**, **Arg3–BENArg3–EIN**, **Arg3–EXT**, **Arg3–FIN**, **Arg3–INS**, **Arg3–ORI**, **Arg4–DES**, **Arg4–EFI**, **ArgL**, **ArgM**,**ArgM–CAU**,**ArgM–EXT**,**ArgM–FIN**,**ArgX**.

For the SC prediction task the features are similar, but not the same. We point out the differences in both directions:

- Features exclusive to the SR system:

  Verb form (**VForm**), verb lemma (**VLemma**), absolute position of the sibling in the clause (**SibAbsPos**), function of the sibling (**SibSynFunc**), preposition of the sibling (**SibPrep**), POS of the second and third words after CW (**CW+2POS, CW+3POS**), feature indicating whether the CW is a named entity (**CWne**, **SIBtot**), syntactic function, syntactic category, preposition and first to fourth word of siblings 8 and 9 (**SynFuncSib8, SynCatSib8, PrepSib8,W1Sib8, W2Sib8, W3Sib8, W4Sib8, SynFuncSib9, SynCatSib9, PrepSib9, W1Sib9, W2Sib9, W3Sib9, W4Sib9**).

- Features exclusive to the SC system:

  AllCats: vector of the syntactic categories of the siblings in the order that they appear in the clause; **AllFuncs**: vector of the functions of the siblings in the order that they appear; **AllFuncsBin** vector with eight binary values that represent if a sibling with that function is present or not; **Sib+1Prep, Sib+2Prep**: prepositions of the two siblings after the verb.

3 Results

3.1 Overall results

SR TASK	PP	Prec.	Recall	$F_{\beta=1}$
Test ca.3LB	74.32%	87.20%	86.52%	86.86
Test ca.CESS	61.62%	83.45%	78.59%	80.95
Overall ca	67.97%	85.32%	82.55%	83.90
Test sp.3LB	68.56%	83.36%	82.85%	83.10
Test sp.CESS	73.98%	85.78%	85.70%	85.74
Overall sp	71.27%	84.57%	84.27%	84.42
Overall SR	69.62%	84.95%	83.41%	84.16

SC TASK	PP	Prec.	Recall	$F_{\beta=1}$
Test ca.3LB	90.86%	90.30%	88.72%	89.50
Test ca.CESS	90.41%	90.20%	88.27%	89.22
Overall ca	90.64%	90.25%	88.50%	89.37
Test sp.3LB	84.12%	80.00%	78.44%	79.21
Test sp.CESS	90.54%	89.89%	89.89%	89.89
Overall sp	86.88%	84.30%	83.36%	83.83
Overall SC	88.67%	87.12%	85.81%	86.46

SRL TASK	PP	Prec.	Recall	$F_{\beta=1}$
Overall ca	–	86.93%	84.49%	85.69
Overall sp	–	84.38%	83.87%	84.12
Overall SRL	–	85.61%	84.17%	84.89

Table 2: *Overall results in the SR (above), SC (middle), and general SRL tasks ('PP': perfect propositions; Prec.: precision; 'ca': Catalan; 'sp': Spanish).*

The overall results of the system are shown in Table 2. The SC system displays a better generalization performance (overall $F_{\beta=1} = 86.46$) than the SR system (overall $F_{\beta=1} = 84.16$), which is also reflected in the average score in terms of correctly identified propositions (88.67% in SC, and 69.62% in SR). The two tasks are inherently different, and there are also marked differences in their example sets. There are less classes in the SC task than in the SR task (cf. Table 3), and they are more homogeneous in the SC task (cf. the entropy rate in Table 3). Additionally, the annotation process might have been different for semantic roles and for verb semantic classes. As for the characteristics of the system, the features selected for the SC task might be more expressive than those selected for the SR task, although the features were selected independently for each task. Additionally, the verbs are easier to identify in the focus selection process because they are marked in the corpus.

	SR task		SC task	
	ca.3LB	sp.3LB	ca.3LB	sp.3LB
Classes	39	48	17	17
Entropy	3.5069	3.6231	2.5609	2.7665

Table 3: *Number of classes and entropy rate in the train corpus (3LB) for Catalan (ca) and Spanish (sp).*

The comparison of results between the 3LB test set and the out-of-domain CESS–ECE set shows that the tendency is different for Spanish and Catalan. The results for Spanish are unexpected because the sp.CESS–ECE test set yields better results: in the SR task, it is processed with $F_{\beta=1}$=85.74, while the sp.3LB is processed with $F_{\beta=1}$=83.10. On the SC task, the sp.CESS–ECE is processed with $F_{\beta=1}$=89.89, while sp.3LB is processed with $F_{\beta=1}$=79.21. The same tendency is observed in the results of the other participants in the task, suggesting that it may be relevant to investigate how the sp.3LB corpus was annotated and partitioned.

The results for Catalan follow the expectations. On the SR task, the $F_{\beta=1}$ rate (80.95) for the out-of-domain ca.CESS–ECE test set is 6 points lower than the $F_{\beta=1}$ rate (86.86) for the ca.3LB test set, and in the SC task, the $F_{\beta=1}$ rate (89.22) for the ca.CESS–ECE test set is also lower than the rate (89.50) for the ca.3LB test set, although the difference is small.

With respect to the robustness of our systems, the results seem to suggest that the SC system is more robust than the SR system. Concerning the difference between the two languages, we observe that the SR system performs better for Spanish (84.42) than for Catalan (83.90), while the SC system performs better for Catalan (89.37) than for Spanish (83.83). The results suggest that the language is not the main factor of the differences in performance, confirming our hypothesis that the task can be approached with the same system for both languages.

3.2 Analysis of the results on the out-of-domain test set

Next, we present detailed results on the Spanish CESS-ECE test (Tables 4 and 5). The differences in score between classes are higher in the SR task than in the SC task. The average of precision and recall is similar in each of the tasks, and both precision and recall are higher for the SC task.

In the SR task class scores are roughly correlated to the frequency of occurrence of classes in the training corpus. Some of the most frequently occurring classes in the test set (Arg0–AGT, Arg1–PAT, Arg1–TEM, Arg2–ATR) are identified at the highest accuracy rates. Aside from the fact that more training examples provide a better chance of being used as nearest neighbors in classification, the feature selection method is also naturally biased towards these classes. High scores attained for medium-frequency classes such as Arg2–BEN can typically be explained by the fact that it has overt markers: in Spanish it is always marked by the Indirect Object function and the prepositions *a* or *para*.

However, some other medium-frequency classes are identified at medium or low accuracy levels of accuracy. For example, the confusion matrix shows that arguments with class ArgM–ADV are assigned ArgM–TMP in 4.35% of cases, ArgM–FIN in 2.89%, ArgM–LOC in 13.04%, ArgM–CAU in 1.44%, Arg2–LOC in 1.44%, and ArgM–MNR 7.24%. Arguments with class ArgM–LOC are assigned ArgM–TMP in 2.53% of cases, Arg2–ATR in 1.26%, Arg0–AGT in 1.26%, ArgM–ADV in 7.59%, ArgM–FIN in 1.26%, ArgM–MNR in 1.26%, and Arg2–LOC in 3,79%. The plausible cause of this is that the selected features are not expressive enough to disambiguate between these confusable outcomes.

In the SC task the three most frequent classes (d2,

SP–CESS	N	Precision	Recall	$F_{\beta=1}$
Overall	1028	85.78%	85.70%	85.74
Arg0–AGT	224	93.21%	91.96%	92.58
Arg0–CAU	6	100%	50%	66.67
Arg1	28	88.46%	82.14%	85.19
Arg1–LOC	1	0.00%	0.00%	0.00
Arg1–PAT	258	93.82%	94.19%	94.00
Arg1–TEM	98	85.71%	91.84%	88.67
Arg2	22	64.29%	81.82%	72.00
Arg2–ATR	73	91.67%	90.41%	91.03
Arg2–BEN	26	100%	100.00%	100.00
Arg2–EFI	3	0.00%	0.00%	0.00
Arg2–EXT	0	0.00%	0.00%	0.00
Arg2–LOC	4	0.00%	0.00%	0.00
Arg2–PAT	1	0.00%	0.00%	0.00
Arg3–ATR	0	0.00%	0.00%	0.00
Arg3–BEN	1	100.00%	100.00%	100.00
Arg3–EIN	0	0.00%	0.00%	0.00
Arg3–FIN	3	100.00%	33.33%	50.00
Arg3–ORI	3	0.00%	0.00%	0.00
Arg4–DES	6	80.00%	66.67%	72.73
ArgL	5	16.67%	20.00%	18.18
ArgM–ADV	69	66.20%	68.12%	67.14
ArgM–CAU	11	62.50%	45.45%	52.63
ArgM–FIN	13	64.71%	84.62%	73.33
ArgM–LOC	79	78.21%	77.22%	77.71
ArgM–MNR	7	40.00%	57.14%	47.06
ArgM–TMP	87	87.65%	81.61%	84.52
V	465	100.00%	100.00%	100.00

Table 4: *Detailed results on the Spanish CESS–ECE test set for the SR task (N: number of appearances in the test corpus).*

SP–CESS	N	Precision	Recall	$F_{\beta=1}$
Overall	465	89.89%	89.89%	89.89
a1	19	85.71%	94.74%	90.00
a2	4	80.00%	100.00%	88.89
b1	9	63.64%	77.78%	70.00
b2	1	100.00%	100.00%	100.00
c1	25	81.82%	72.00%	76.60
c3	57	84.85%	98.25%	91.06
c5	3	75.00%	100.00%	85.71
d1	14	78.57%	78.57%	78.57
d2	248	97.00%	91.13%	93.97
d3	78	91.78%	85.90%	88.74
d4	1	14.29%	100.00%	25.00
d5	1	33.33%	100.00%	50.00
e1	5	100.00%	100.00%	100.00

Table 5: *Detailed results on the Spanish CESS–ECE test set for the SC task (N:number of appearances in the test corpus).*

d3, c3) are predicted at high levels of accuracy. At the same time, some of the less frequent classes also receive high scores (a2, b2, c5, e1). In contrast with the SR task, all classes receive a non-zero score.

3.3 Analysis of the results for all semantic roles

Table 6 shows the $F_{\beta=1}$ rates for all individual semantic roles in the test sets. Most of the large differences between scores obtained for the same semantic role in different test sets can be explained by the fact that these semantic roles have a low frequency (Arg0–EXP, Arg1–EXT, Arg1–LOC, Arg2–EFI, Arg2–EXT, Arg2–LOC, Arg3–BEN, Arg3–FIN, Arg3–ORI, ArgL, ArgM–MNR). Some semantic roles are stable across test sets and receive a medium score (Arg0–AGT, Arg1, Arg1–PAT, Arg1–TEM, Arg2, Arg2–ATR, Arg2–BEN). This might mean that these semantic roles are frequent, that the features are expressive for these classes, and possibly that they are annotated consistently.

At the same time, some roles receive very different scores in the different test sets (Arg2–LOC, Arg2–DES, ArgM–CAU, ArgL, ArgM–MNR, ArgM–TMP). This might be caused by different frequencies of the semantic roles in the corpus, but also by inconsistent annotation.

	ca.3LB	ca.CESS	sp.3LB	sp.CESS
Arg0–AGT	93.21	91.47	90.79	92.58
Arg0–CAU	40.00	42.11	45.45	66.67
Arg0–EXP	–	0.00	50.00	–
Arg0–TEM	–	–	0.00	–
Arg1	75.68	80.00	79.17	85.19
Arg1–AGT	–	–	0.00	–
Arg1–EXT	0.00	–	100.00	–
Arg1–LOC	0.00	66.67	0.00	0.00
Arg1–PAT	94.50	93.46	92.17	94.00
Arg1–TEM	90.99	88.57	89.95	88.67
Arg2	78.38	77.14	74.07	72.00
Arg2–ATR	92.77	92.72	95.38	91.03
Arg2–BEN	100.00	100.00	94.74	100.00
Arg2–EFI	–	–	40.00	0.00
Arg2–EXT	66.67	40.00	–	0.00
Arg2–LOC	36.36	57.14	30.43	0.00
Arg2–ORI	–	0.00	–	–
Arg2–PAT	–	–	–	0.00
Arg3–ATR	0.00	0.00	0.00	–
Arg3–BEN	–	–	0.00	100.00
Arg3–EIN	0.00	0.00	–	0.00
Arg3–FIN	100.00	0.00	0.00	50.00
Arg3–ORI	25.00	0.00	61.54	0.00
Arg4–DES	72.73	54.55	50.00	72.73
Arg4–EFI	0.00	0.00	–	–
ArgL	80.00	33.33	20.00	18.18
ArgM	–	0.00	–	–
ArgM–ADV	63.79	61.07	71.89	67.14
ArgM–CAU	80.95	66.67	78.79	52.63
ArgM–EXT	0.00	0.00		–
ArgM–FIN	84.00	84.24	87.80	73.33
ArgM–LOC	70.31	75.24	70.24	77.71
ArgM–MNR	51.16	21.05	53.66	47.06
ArgM–PAT	–	–	0.00	–
ArgM–TMP	91.43	41.48	77.46	84.52
V	99.22	99.25	99.10	100.00

Table 6: $F_{\beta=1}$ *rate for all semantic roles in the four test sets.*

3.4 Analysis of features selected for SR

Table 7 shows the twenty features with the highest gain ratio in the SR task for Catalan and Spanish. The feature SibSynFunc has the highest gain ratio (Catalan 0.7198, Spanish 0.7661). Among these twenty features, sixteen are the same in both languages. The four features exclusive to Catalan are Arg2–INS, Arg0–EXP,

Arg0–TEM, and CWPOSType. Mostly these are features from the verb lexicon. For Spanish the deviating features are also from the verb lexicon: Arg2–ADV, Arg2–ORI, Arg0, and Arg3–BEN.

ca.roles		sp.roles	
feat.	GR	feat.	GR
SybSynFunc	0.7198	SybSynFunc	0.7661
Arg2–INS	0.4449	SibSynCat	0.4128
SibPrep	0.4179	SibPrep	0.4124
SibSynCat	0.4069	Arg2–ADV	0.3834
ATRRelPos	0.3745	SibRelPos	0.3554
SibRelPos	0.3444	ATRRelPos	0.3451
Arg0–EXP	0.3428	SibPOSW1	0.3224
SibPOSW1	0.3369	Arg3–FIN	0.3065
Arg3–FIN	0.3363	SibLemW1	0.2927
CWPOS	0.3095	Arg2–ORI	0.2871
SibLemW1	0.3082	CWPOS	0.2853
Arg1–PAT	0.3035	Arg0	0.2729
CREGRelPos	0.2653	CWtmp	0.2548
Arg0–TEM	0.2644	Arg1–PAT	0.2474
CIRelPos	0.2481	CWord	0.2466
Arg1–TEM	0.2444	CREGRelPos	0.2428
CWord	0.2422	CIRelPos	0.2372
CWtmp	0.2402	Arg1–TEM	0.2367
CWPOStype	0.2399	Arg3–BEN	0.2367
CWLemma	0.2377	CWLemma	0.2341

Table 7: *Features with the highest Gain Ratio in the SR task.*

To sum up, features do not obtain the same gain ratio for both languages, but they show the same tendency. The top features encode information about the syntactic function, the preposition, the syntactic category, and the relative position of the focus sibling; the lemma and POS of the first word of the current sibling; the POS and word of the content word; the relative position of the sibling with function ATR, CREG and CI; and information from the verb lexicon.

3.5 Analysis of features selected for SC

Table 8 shows the twenty features with the highest gain ratio in the SC task for Catalan and Spanish. Most of the features originate from the verb lexicon. The feature with the highest gain ratio in Catalan is Arg0–EXP (0.7328), whereas in Spanish it is ATRRelPOS (0.5515).

A comparison of both systems shows that in the SC system the features with the highest gain ratio are mostly features from the verb lexicon, whereas in the SR system only some features from the lexicon are the top positions. The features CWLemma, ATRRelPos, and CWLemma are in the top positions in both systems, as well as the lexicon features Arg0–EXP, Arg1–PAT, and Arg1–TEM.

3.6 Analysis of the effect of removing features

Tables 9 and 10 contain information about the effects of removing features from the SR system. Table 9 focuses on the effects of removing groups of features. Removing the features that provide information

ca.verbs		sp.verbs	
feat.	GR	feat.	GR
Arg0–EXP	0.7328	ATRRelPos	0.5515
ATRRelPos	0.5470	Arg3–FIN	0.4964
Arg1–TEM	0.4397	Arg3–BEN	0.4704
Arg2–EXT	0.3929	Arg1–TEM	0.4294
Arg0–AGT	0.3849	Arg0–EXP	0.3655
Arg2–LOC	0.3302	Arg2–BEN	0.3530
Arg1–PAT	0.3248	Arg0–AGT	0.3385
Arg2–BEN	0.3215	Arg2–ATR	0.3334
CIRelPos	0.3176	Arg0–CAU	0.3333
ArgX	0.3131	Arg3	0.3330
Arg3–ATR	0.2982	Arg0	0.3329
Arg4–EFI	0.2980	Arg1–PAT	0.3291
Arg3–EIN	0.2940	Arg2	0.3046
Arg2–ATR	0.2918	VCau	0.3006
Arg2–INS	0.2915	CIRelPos	0.2915
Arg2–EFI	0.2815	Arg2–EFI	0.2850
Arg2	0.2770	Arg1–EXT	0.2824
Arg3-BEN	0.2713	Arg2–PAT	0.2798
CWLemma	0.2565	CWLemma	0.2658
Arg3–EXT	0.2558	SibLemW1	0.2369

Table 8: *Features with the highest Gain Ratio in the SC task.*

about the sibling in focus ('Sibling') causes a clear decrease in the system's performance (on average 21.85 points of F-score). Removing the verb lexicon features ('Lexicon Roles') and the features of the verb in focus ('Verb') also causes a decrease in the system's performance, but much lower. Removing the features of the clause containing the verb in focus ('Clause') causes a slight decrease, and removing the features that describe properties of the content word ('CW')causes different effects in each test set, but just a slight decrease or increase. These results show that the most expressive features in this task are the features on the sibling in focus.

	ca.3LB	ca.CESS	sp.3LB	sp.CESS
With all	86.25	80.25	82.80	85.84
- Sibling	**-18.42**	**-19.32**	**-24.11**	**-25.58**
- Lexicon Roles	-1.66	-2.51	-3.57	-2.43
- Verb	-2.01	-3.59	-2.51	-1.17
- Clause	-0.70	-1.62	-0.62	-0.39
- CW	+0.74	-0.17	-0.32	+0.19

Table 9: *Effect of removing groups of features from the SR system. (Overall $F_{\beta=1}$).*

Table 10 contains information about the effects of removing the ten individual features that have the highest gain ratio in the Catalan and Spanish training corpora (listed in Table 7). As expected, removing the feature SybSynFunc causes a clear decrease in the results (on average 5.18 points of F-score).

	ca.3LB	ca.CESS	sp.3LB	sp.CESS
With all	86.25	80.25	82.80	85.84
- SybSynFunc	**-5.76**	**-2.47**	**-6.37**	**-5.90**
- SibPrep	-0.44	-0.33	-0.62	-1.13
- Arg0–EXP	0.00	0.00	-0.16	-0.10
- Arg3–FIN	0.00	0.00	0.00	0.00
- Arg2–ADV	0.00	0.00	0.00	0.00
- Arg2–ORI	0.00	0.00	0.00	0.00
- Arg2–INS	0.00	0.00	-0.08	+0.10
- SibPOSW1	0.00	-0.08	-0.47	+0.10
- SibRelPos	+0.18	-0.37	-0.39	-0.34
- SibLemW1	+0.26	-0.33	-0.08	0.00
- SibSynCat	+0.18	-0.29	-0.06	+0.04
- CWPOS	-0.17	+0.18	-0.08	+0.10
- ATRRelPos	0.00	+0.30	+0.07	+0.10
- 20 feats. with highest GR	-1.75	-1.88	-2.42	-2.28

Table 10: *Effect of removing features with high gain ratio from the SR system (overall $F_{\beta=1}$).*

4 Conclusions

We presented a memory-based semantic role labeling (SRL) system for Catalan and Spanish that makes use of full syntactic information. We approached the general SRL task as two distinct classification problems: the assignment of semantic roles to arguments of verbs, and the assignment of semantic classes to verbs. Building on a pool of features from which we selected subsets appropriate to each subtask, we trained two similar classifiers on the two subtasks using the IB1 classifier as implemented in TiMBL (version 5.1) [7]. We reported an overall performance of the system of 85.69 $F_{\beta=1}$ for Catalan, and 84.12 $F_{\beta=1}$ for Spanish.

The results show that a uniform single-classifier system can produce competitive results in both tasks. It performs slightly better on the SC task, which might be caused by several reasons: apart from the fact that the tasks are inherently different and SC may simply be easier, there are less classes in the SC task than in the SR task, with stronger predictability from the same pool of features. Other factors such as the consistency of the annotation might play a role. Additionally, the features selected for the SC task might be more expressive than those selected for the SR task. Results also show that the two problems can be solved in largely the same way for both languages. On the SC task the approach results in higher generalization performance for Catalan, and on the SR task the Spanish system is better. Finally, the effects of removing groups of features show that the most expressive features in the SR task are clearly the features that provide information about the sibling in focus.

Acknowledgements

This research has been funded by the postdoctoral grant EX2005–1145 awarded by the Ministerio de Educación y Ciencia of Spain to the project *Técnicas semi-automáticas para el etiquetado de roles semánticos en corpus del español*. We are grateful to Bertjan Busser for his contribution to programming the system and to the three anonymous reviewers.

References

[1] X. Carreras and L. Màrquez. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL–2004*, Boston MA, USA, 2004.

[2] X. Carreras and L. Màrquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL–2005*, Ann Arbor, Michigan, June 2005.

[3] S. Cost and S. Salzberg. A weighted nearest neighbour algorithm for learning with symbolic features. *Machine Learning*, 10:57–78, 1993.

[4] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13:21–27, 1967.

[5] W. Daelemans and A. van den Bosch. *Memory-based language processing.* Cambridge University Press, Cambridge, UK, 2005.

[6] W. Daelemans, A. Van den Bosch, and J. Zavrel. Forgetting exceptions is harmful in language learning. *Machine Learning, Special issue on Natural Language Learning*, 34:11–41, 1999.

[7] W. Daelemans, J. Zavrel, K. V. der Sloot, and A. V. den Bosch. TiMBL: Tilburg memory based learner, version 5.1, reference guide. Technical Report Series 04-02, ILK, Tilburg, The Netherlands, 2004.

[8] B. Decadt, V. Hoste, W. Daelemans, and A. Van den Bosch. GAMBL, genetic algorithm optimization of memory-based WSD. In R. Mihalcea and P. Edmonds, editors, *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 108–112, New Brunswick, NJ, 2004. ACL.

[9] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[10] V. Hoste, I. Hendrickx, W. Daelemans, and A. Van den Bosch. Parameter optimization for machine learning of word sense disambiguation. *Natural Language Engineering*, 8(4):311–325, 2002.

[11] M. Komachi, Y. Matsumoto, and M. Nagata. Phrase reordering for statistical machine translation based on predicate-argument structure. In *Proceedings of the International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation*, pages 77–82, Kyoto, Japan, 2006.

[12] L. Màrquez, P. Comas, J. Giménez, and N. Català. Semantic role labeling as sequential tagging. In *Proceedings of CoNLL–2005*, Ann Arbor, Michigan, 2005.

[13] L. Màrquez, L. Villarejo, M. Martí, and M. Taulé. Semeval-2007 task 09: Multilevel semantic annotation of Catalan and Spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 42–47, 2007.

[14] R. Morante and B. Busser. ILK2: Semantic role labelling for Catalan and Spanish using TiMBL. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 183–186, 2007.

[15] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. Using predicate–argument structures for information extraction. In *Proceedings of the ACL 2003*, 2003.

[16] E. Tjong Kim Sang, S. Canisius, A. van den Bosch, and T. Bogers. Applying spelling error correction techniques for improving semantic role labelling. In *Proceedings of CoNLL-2005*, pages 229–232, Ann Arbor, Michigan, 2005.

[17] K. Toutanova, A. Haghighi, and C. Manning. Joint learning improves semantic role labeling. In *Proceedings of ACL–05*, Ann Arbor, Michigan, 2005.

[18] N. Xue and M. Palmer. Calibrating features for semantic role labeling. In *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.