ILK2: Semantic Role Labelling for Catalan and Spanish using TiMBL

Roser Morante, Bertjan Busser
ILK, Dept. of Language and Information Sciences
Tilburg University, P.O.Box 90153
NL-5000 LE Tilburg, The Netherlands
{R.Morante,G.J.Busser}@uvt.nl

Abstract

In this paper we present a semantic role labeling system submitted to the task *Multilevel Semantic Annotation of Catalan and Spanish* in the context of SemEval–2007. The core of the system is a memory–based classifier that makes use of full syntactic information. Building on standard features, we train two classifiers to predict separately the semantic class of the verb and the semantic roles.

1 Introduction

Semantic role labelling (SRL) has been addressed in the CoNLL–2004 and CoNLL–2005 Shared Tasks (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005) for English. In the task *Multilevel Semantic Annotation of Catalan and Spanish* of the SemEval competition 2007, the target are two different languages. The general SRL task consists of two tasks: prediction of semantic roles (SR) and prediction of the semantic class of the verb (SC).

The data provided in the task (Màrquez et al., 2007) are sentences annotated with lemma, POS tags, syntactic information, semantic roles, and the semantic classes of the verb. A training corpus for Catalan (ca.3LB) and another for Spanish (sp.3LB) are provided. Although the setting is similar to the CoNLL–Shared Task 2005, three relevant differences are that the corpora are significantly smaller, that the syntactic information is based on a manually corrected treebank, which contains also syntactic functions (i.e. direct object, indirect object, etc.), and that the set of semantic roles is larger, especially for core arguments.

Our goal is to check whether simple individual systems could produce competitive results in both subtasks, and whether they would be robust enough when applied to two languages and to the held–out test sets provided.

2 System description

We approach the SRL task as two classification problems: prediction of SR and prediction of SC. We hypothesize that the two problems can be solved in the same way for both languages. We build two very similar systems that differ only in some of the features used, as we explain below.

The task is solved in three phases: 1) A pre–processing phase that is very similar to the sequentialization in (Màrquez et al., 2005). We call it *focus selection*. It consists of identifying the potential candidates to be assigned a semantic role or a semantic verb class. 2) The classification. 3) Some limited postprocessing.

2.1 Focus selection

The system starts by finding the target verb (which is marked in the corpus as such). Then, it finds the complete form of the verb (that 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 are under the same clause. Our assumption is that all siblings of the verb are potential candidates for semantic roles. The focus selection process produces two groups of focus tokens: on the one hand, the verbs and, on the other, the siblings of

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 183–186,
Prague, June 2007. ©2007 Association for Computational Linguistics

the verbs. These tokens will be the instances in each training set. Table 1 shows the number of training and test instances for each subtask.

	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).

2.2 Classification

In both systems we approach the classification task in one step, predicting directly the SR and the SC class. This means that in the SR task we do not perform a previous classification to select the tokens that might be assigned a role. We assume that all verbs belong to a class. As for the SR, 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. 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.

Regarding the **learning algorithm**, we use the IB1 classifier as implemented in TiMBL (version 5.1) (Daelemans et al., 2004), a supervised inductive algorithm for learning classification tasks based on the k nearest neighbor (k-nn) algorithm. In IB1, similarity is defined by a feature–level distance metric between a test instance and a memorized training instance. The metric combines a per–feature value–based distance metric with global feature weights that account for relative differences in importance of the features.

The TiMBL parameters used in the systems are the IB1 algorithm, the Jeffrey Divergence as feature metric, MVDM threshold at level 1, weighting using GainRatio, k=11, and weighting neighbors as function of their Inverse Linear Distance (for details we refer the reader to the TiMBL reference guide (Daelemans et al., 2004)).

As for the **features**, we started by using the same feature set for both classifiers and then, after some experimentation, we decided to use slightly different feature sets for the two sub-tasks. Most of the features we designed are features that have become standard for the SRL task (Gildea and Jurafsky, 2002; Xue and Palmer, 2004; Carreras and Màrquez, 2004; Carreras and Màrquez, 2005). In our system, the features relate to the verb, the verb siblings, what we take to be the content word of the siblings, the clause, and the relation verb–arguments. Additionally, we added lexical features extracted from the verb lexicon provided for the task, and from WordNet.

After experimenting with 323 features, we selected 98 for the SR task and 77 for the SC subclass. 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 10-fold cross-validation experiments; 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 also performed experiments applying the feature selection process reported in (Tjong Kim Sang et al., 2005), 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 subtask are the following:

• Features on the verb (6). They are shared by all the instances that represent phrases belonging to the same clause:

VForm; **VLemma**; **VCau**: binary feature 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 on the sibling in focus (12):

SibSynCat: syntactic category; **SibSynFunc**: syntactic function; **SibPrep**: preposition; **SibLemW1, SibPOSW1, SibLemW2, SibPOSW2, SibLemW3, SibPOSW3**: lemma and POS tag 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 the properties of the content word (CW) of the focus sibling (13): in the case of prepositional phrases the CW is the head of the first noun phrase; in cases of coordination, we only take the first element of the coordination.

CWord; **CWLemma**; **CWPOS**: we take only the first character of the POS tags provided; **CWPOSType**: the type of POS, second character of the POS tags provided; **CWGender**; **CWne**: binary feature that indicates if the CW is a named entity; **CWtmp, CWloc**: binary 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.

CWwnsc1, CWwnsc2, CWwnsc3: additionally, if the CW is a noun, we extract information from WordNet (Fellbaum, 1998) about the first, second, and third more frequent semantic classes of the CW in WordNet. We cannot decide on a single one because the corpus is not disambiguated. The semantic class corresponds to the lexicographer files in WN3.0. For nouns there are 25 file numbers.

• Features on the clause (24):

CCtot: total number of siblings with function CC (circumstancial complement); **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**: binary 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 lexicon of verbal frames (43) that the task organizers provided. We access the lexicon to check if it is possible for a verb to have a certain semantic role. We check it for all semantic role classes, except for ArgX-Ag, ArgX-Cau, ArgX-Pat, ArgX-Tem because they proved not to be informative. The features are binary.

For the SC prediction task the features are similar, but not exactly the same. Both systems contain some features about all candidate arguments. We point out the differences:

• Features that are in the SR system and that are not in the SC 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 tag of the second and third words after CW (**CW+2POS, CW+3POS**), information about the WN classes of the CW (**CWwnsc1, CWwnsc2, CWwnsc3**), feature about the CW being 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 that are only in 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; **AllFuncs-Bin** 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.

2.3 Postprocessing

As for the **postprocessing phase**, it consists of six simple rules to correct some basic errors in predicting some types of ArgM arguments. It only applies to the SR task. The rules are the following ones:

1. If prediction = ArgM–LOC, ArgM–MNR or ArgM–ADV, and either {**SibPrep** = 'durante' or 'durant'}, or {**SibSynCat** = sn and one of the WN semantic classes = 28}, then prediction = ArgM-TMP.

2. If prediction = ArgM–LOC, ArgM–MNR or ArgM–ADV, and **CWLemma** is a temporal adverb, then prediction = ArgM–TMP.

3. If prediction = ArgM–TMP and one of the WN classes = 15, then prediction = ArgM–LOC.

4. If prediction = ArgM–TMP, ArgM-MNR or ArgM-ADV, and **CWLemma** = locative adverb, then prediction = ArgM-LOC.

5. If prediction = ArgM-TMP or ArgM-ADV, and **CWwnsc1** = 15, and **SibPrep** = 'en' or 'desde' or 'hacia' or 'a' or 'des_de' or 'cap_a', then prediction = ArgM–LOC.

6. If prediction = ArgM–ADV and **CWLemma** = causal conjunction, then prediction = ArgM–CAU.

We are aware of the fact that these are very simple rules and that more elaborate postprocessing techniques can be applied, like the ones used in (Tjong Kim Sang et al., 2005) in order to make sure that the same role was not predicted more than once in the same clause.

SR TASK	Perf.Props	Precision	Recall	$F_{\beta=1}$
Test ca.3LB	73.35%	86.59%	85.91%	86.25
Test ca.CESS	60.55%	82.60%	78.03%	80.25
Overall ca	67.24%	84.72%	82.12%	83.40
Test sp.3LB	68.07%	83.05%	82.54%	82.80
Test sp.CESS	73.76%	85.88%	85.80%	85.84
Overall sp	70.52%	84.30%	83.98%	84.14
Overall SR	68.96%	84.50%	83.07%	83.78

SC TASK	Perf.Props	Precision	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	Perf.Props	Precision	Recall	$F_{\beta=1}$
Overall ca	–	86.44%	84.08 %	85.24
Overall sp	–	84.30%	83.78 %	84.04
Overall SRL	–	85.32%	83.93 %	84.62

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

3 Results

The overall official results of the system are shown in Table 2. The SC system performs better (overall F_1 = 86.46) than the SR system (overall F_1 = 83.78). In global, the systems perform better for Catalan (overall F_1 = 85.24) than for Spanish (overall F_1 = 84.04), although the SC system performs better for Catalan (89.37 vs. 86.46), and the SR system performs better for Spanish (84.14 vs 83.40).

Striking results are that the SR system gets significantly better results with the held–out test for Spanish, and that both of the complete SRL systems get significantly better results with the held–out test for Spanish. This might be due to differences in the process of gathering and annotation of the corpus.

SP–CESS	F	Precision	Recall	$F_{\beta=1}$
Overall		85.88%	85.80%	85.84
Arg0–AGT	16.19%	92.83%	92.41%	92.62
Arg0–CAU	1.23%	100%	50%	66.67
Arg1	1.79%	88.46%	82.14%	85.19
Arg1–LOC	0.11%	0.00%	0.00%	0.00
Arg1–PAT	20.09%	93.82%	94.19%	94.00
Arg1–TEM	14.08%	86.54%	91.84%	89.11
Arg2	2.05%	68.00%	77.27%	72.34
Arg2–ATR	9.88%	91.67%	90.41%	91.03
Arg2–BEN	2.40%	96.30%	100.00%	98.11
Arg2–EFI	0.19%	0.00%	0.00%	0.00
Arg2–EXT	0.19%	0.00%	0.00%	0.00
Arg2–LOC	1.13%	0.00%	0.00%	0.00
Arg2–PAT	0.01%	0.00%	0.00%	0.00
Arg3–ATR	0.05%	0.00%	0.00%	0.00
Arg3–BEN	0.16%	100.00%	100.00%	100.00
Arg3–EIN	0.08%	0.00%	0.00%	0.00
Arg3–FIN	0.04%	100.00%	33.33%	50.00
Arg3–ORI	0.29%	0.00%	0.00%	0.00
Arg4–DES	0.60%	83.33%	83.33%	83.33
ArgL	0.71%	16.67%	20.00%	18.18
ArgM–ADV	10.67%	68.12%	68.12%	68.12
ArgM–CAU	1.50%	55.56%	45.45%	50.00
ArgM–FIN	1.30%	64.71%	84.62%	73.33
ArgM–LOC	4.94%	78.21%	77.22%	77.71
ArgM–MNR	2.28%	36.36%	57.14%	44.44
ArgM–TMP	7.19%	88.75%	81.61%	85.03
V	–	100.00%	100.00%	100.00

Table 3: Detailed results on the Spanish CESS–ECE test corpus for the SR subtask. F: frequency of the semantic roles in the training corpus, without counting V.

Table 3 shows detailed results on the Spanish CESS–ECE corpus for the SR task. Low scores are generally related to low frequency of the SR in the training corpus, and high scores are related to high frequency or to overt marking of the SR.

4 Conclusions

We have presented two memory–based SRL systems that make use of full syntactic information and approach the tasks in three steps. Results show that rather simple individual systems can produce competitive results in both tasks, and that they are robust enough to be applied to two languages and to the held–out test sets provided. Improvements of the systems would consist in improving the focus selection step, and applying more elaborate techniques for feature selection and postprocessing.

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 semiautomáticas para el etiquetado de roles semánticos en corpus del español*. We would like to thank Martin Reynaert, Caroline Sporleder, Antal van den Bosch, and the anonymous reviewers for their comments and suggestions.

References

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

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

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

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

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

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

Ll. Màrquez, M.A. Martí, M. Taulé, and L. Villarejo. 2007. SemEval-2007 Task 09: Multilevel semantic annotation of catalan and spanish. In *Proceedings of SemEval-2007, the 4th Workshop on Semantic Evaluations*, Prague, Czech Republic.

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

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