Generalizing over Lexical Features: Selectional Preferences for Semantic Role Classification

Beñat Zapirain, Eneko Agirre
Ixa Taldea
University of the Basque Country
Donostia, Basque Country
{benat.zapirain,e.agirre}@ehu.es

Lluís Màrquez
TALP Research Center
Technical University of Catalonia
Barcelona, Catalonia
lluism@lsi.upc.edu

Abstract

This paper explores methods to alleviate the effect of lexical sparseness in the classification of verbal arguments. We show how automatically generated selectional preferences are able to generalize and perform better than lexical features in a large dataset for semantic role classification. The best results are obtained with a novel second-order distributional similarity measure, and the positive effect is specially relevant for out-of-domain data. Our findings suggest that selectional preferences have potential for improving a full system for Semantic Role Labeling.

1 Introduction

Semantic Role Labeling (SRL) systems usually approach the problem as a sequence of two subtasks: argument *identification* and *classification*. While the former is mostly a syntactic task, the latter requires semantic knowledge to be taken into account. Current systems capture semantics through lexicalized features on the predicate and the head word of the argument to be classified. Since lexical features tend to be sparse (especially when the training corpus is small) SRL systems are prone to overfit the training data and generalize poorly to new corpora.

This work explores the usefulness of selectional preferences to alleviate the lexical dependence of SRL systems. Selectional preferences introduce semantic generalizations on the type of arguments preferred by the predicates. Therefore, they are expected to improve generalization on infrequent and unknown words, and increase the discriminative power of the argument classifiers.

For instance, consider these two sentences:

JFK was assassinated (in Dallas)$_{Location}$
JFK was assassinated (in November)$_{Temporal}$

Both share syntactic and argument structure, so the lexical features (i.e., the words 'Dallas' and 'November') represent the most important knowledge to discriminate between the two different adjunct roles. The problem is that, in new text, one may encounter similar expressions with new words like Texas or Autumn.

We propose a concrete classification problem as our main evaluation setting for the acquired selectional preferences: *given a verb occurrence and a nominal head word of a constituent dependant on that verb, assign the most plausible role to the head word according to the selectional preference model*. This problem is directly connected to argument classification in SRL, but we have isolated the evaluation from the complete SRL task. This first step allows us to analyze the potential of selectional preferences as a source of semantic knowledge for discriminating among different role labels. Ongoing work is devoted to the integration of selectional preference–derived features in a complete SRL system.

2 Related Work

Automatic acquisition of selectional preferences is a relatively old topic, and will mention the most relevant references. Resnik (1993) proposed to model selectional preferences using semantic classes from WordNet in order to tackle ambiguity issues in syntax (noun-compounds, coordination, PP-attachment).

Brockman and Lapata (2003) compared several *class-based* models (including Resnik's selectional preferences) on a syntactic plausibility judgement task for German. The models return weights for (verb, syntactic_function, noun) triples, and the correlation with human plausibility judgement is used for evaluation. Resnik's selectional preference scored best among class-based methods, but it performed equal to a simple, purely lexical, conditional probability model.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 73–76,
Suntec, Singapore, 4 August 2009.

Distributional similarity has also been used to tackle syntactic ambiguity. Pantel and Lin (2000) obtained very good results using the distributional similarity measure defined by Lin (1998).

The application of selectional preferences to semantic roles (as opposed to syntactic functions) is more recent. Gildea and Jurafsky (2002) is the only one applying selectional preferences in a real SRL task. They used distributional clustering and WordNet-based techniques on a SRL task on FrameNet roles. They report a very small improvement of the overall performance when using distributional clustering techniques. In this paper we present complementary experiments, with a different role set and annotated corpus (PropBank), a wider range of selectional preference models, and the analysis of out-of-domain results.

Other papers applying semantic preferences in the context of semantic roles, rely on the evaluation on pseudo tasks or human plausibility judgments. In (Erk, 2007) a distributional similarity–based model for selectional preferences is introduced, reminiscent of that of Pantel and Lin (2000). The results over 100 frame-specific roles showed that distributional similarities get smaller error rates than Resnik and EM, with Lin's formula having the smallest error rate. Moreover, coverage of distributional similarities and Resnik are rather low. Our distributional model for selectional preferences follows her formalization.

Currently, there are several models of distributional similarity that could be used for selectional preferences. More recently, Padó and Lapata (2007) presented a study of several parameters that define a broad family of distributional similarity models, including publicly available software.

Our paper tests similar techniques to those presented above, but we evaluate selectional preference models in a setting directly related to SR classification, i.e., given a selectional preference model for a verb we find the role which fits best for a given head word. The problem is indeed qualitatively different: we do not have to choose among the head words competing for a role (as in the papers above) but among selectional preferences competing for a head word.

3 Selectional Preference Models

In this section we present all the variants for acquiring selectional preferences used in our study, and how we apply them to the SR classification.

WordNet-based SP models: we use Resnik's selectional preference model.

Distributional SP models: Given the availability of publicly available resources for distributional similarity, we used 1) a ready-made thesaurus (Lin, 1998), and 2) software (Padó and Lapata, 2007) which we run on the British National Corpus (BNC).

In the first case, Lin constructed his thesaurus based on his own similarity formula run over a large parsed corpus comprising journalism texts. The thesaurus lists, for each word, the most similar words, with their weight. In order to get the similarity for two words, we could check the entry in the thesaurus for either word. But given that the thesaurus is not symmetric, we take the average of both similarities. We will refer to this similarity measure as sim^{th}_{lin}. Another option is to use second-order similarity, where we compute the similarity of two words using the entries in the thesaurus, either using the cosine or Jaccard measures. We will refer to these similarity measures as sim^{th2}_{jac} and sim^{th2}_{cos} hereinafter.

For the second case, we tried the optimal parameters as described in (Padó and Lapata, 2007, p. 179): word-based space, medium context, log-likelihood association, and 2,000 basis elements. We tested Jaccard, cosine and Lin's measure (Lin, 1998) for similarity, yielding sim_{jac}, sim_{cos} and sim_{lin}, respectively.

3.1 Role Classification with SP Models

Given a target sentence where a predicate and several potential argument and adjunct head words occur, the goal is to assign a role label to each of the head words. The classification of candidate head words is performed independently of each other.

Since we want to evaluate the ability of selectional preference models to discriminate among different roles, this is the only knowledge that will be used to perform classification (avoiding the inclusion of any other feature commonly used in SRL). Thus, for each head word, we will simply select the role (r) of the predicate (p) which fits best the head word (w). This selection rule is formalized as:

$R(p, w) = \arg\max_{r \in Roles(p)} S(p, r, w)$

being $S(p, r, w)$ the prediction of the selectional preference model, which can be instantiated with all the variants mentioned above.

For the sake of comparison we also define a lexical baseline model, which will determine the contribution of lexical features in argument classification. For a test pair (p, w) the model returns the role under which the head word occurred most often in the training data given the predicate.

4 Experimental Setting

The data used in this work is the benchmark corpus provided by the CoNLL-2005 shared task on SRL (Carreras and Màrquez, 2005). The dataset, of over 1 million tokens, comprises PropBank sections 02-21 for training, and sections 24 and 23 for development and test, respectively. In these experiments, NEG, DIS and MOD arguments have been discarded because, apart from not being considered "pure" adjunct roles, the selectional preferences implemented in this study are not able to deal with non-nominal argument heads.

The predicate–rol–head (p, r, w) triples for generalizing the selectional preferences are extracted from the arguments of the training set, yielding 71,240 triples, from which 5,587 different predicate-role selectional preferences (p, r) are derived by instantiating the different models in Section 3.

Selectional preferences are then used, to predict the corresponding roles of the (p, w) pairs from the test corpora. The test set contains 4,134 pairs (covering 505 different predicates) to be classified into the appropriate role label. In order to study the behavior on out-of-domain data, we also tested on the PropBanked part of the Brown corpus. This corpus contains 2,932 (p, w) pairs covering 491 different predicates.

The performance of each selectional preference model is evaluated by calculating the standard *precision*, *recall* and F_1 measures. It is worth mentioning that none of the models is able to predict the role when facing an unknown head word. This happens more often with WordNet based models, which have a lower word coverage compared to distributional similarity–based models.

5 Results and Discussion

The results are presented in Table 1. The lexical row corresponds to the baseline lexical match method. The following row corresponds to the WordNet-based selectional preference model. The distributional models follow, including the results obtained by the three similarity formulas on the co-occurrences extracted from the BNC (sim_{Jac}, sim_{cos} sim_{Lin}), and the results obtained when using Lin's thesaurus directly (sim^{th}_{Lin}) and as a second-order vector (sim^{th2}_{Jac} and sim^{th2}_{cos}).

	prec.	rec.	F_1	prec.	recall	F_1
$lexical$	**.779**	.349	.482	**.663**	.059	.108
res	.589	.495	.537	.505	.379	.433
sim_{Jac}	.573	.564	.569	.481	.452	.466
sim_{cos}	.607	.598	.602	.507	.476	.491
sim_{Lin}	.580	.560	.570	.500	.470	.485
sim^{th}_{Lin}	.635	.625	.630	.494	.464	.478
sim^{th2}_{Jac}	.657	**.646**	**.651**	.531	**.499**	**.515**
sim^{th2}_{cos}	.654	.644	.649	.531	**.499**	**.515**

Table 1: Results for WSJ test (left), and Brown test (right)

As expected, the lexical baseline attains very high precision in all datasets, which underscores the importance of the lexical head word features in argument classification. The recall is quite low, specially in Brown, confirming and extending (Pradhan et al., 2008), which also reports similar performance drops when doing argument classification on out-of-domain data.

One of the main goals of our experiments is to overcome the data sparseness of lexical features both on in-domain and out-of-domain data. All our selectional preference models improve over the lexical matching baseline in recall, up to 30 absolute percentage points in the WSJ test dataset and 44 absolute percentage points in the Brown corpus. This comes at the cost of reduced precision, but the overall F-score shows that all selectional preference models improve over the baseline, with up to 17 absolute percentage points on the WSJ datasets and 41 absolute percentage points on the Brown dataset. The results, thus, show that selectional preferences are indeed alleviating the lexical sparseness problem.

As an example, consider the following head words of potential arguments of the verb *wear* found in the test set: *doctor*, *men*, *tie*, *shoe*. None of these nouns occurred as heads of arguments of *wear* in the training data, and thus the lexical feature would be unable to predict any role for them. Using selectional preferences, we successfully assigned the `Arg0` role to *doctor* and *men*, and the `Arg1` role to *tie* and *shoe*.

Regarding the selectional preference variants, WordNet-based and first-order distributional similarity models attain similar levels of precision, but the former are clearly worse on recall and F_1.

The performance loss on recall can be explained by the worse lexical coverage of WordNet when compared to automatically generated thesauri. Examples of words missing in WordNet include abbreviations (e.g., *Inc.*, *Corp.*) and brand names (e.g., *Texaco*, *Sony*). The second-order distributional similarity measures perform best overall, both in precision and recall. As far as we know, it is the first time that these models are applied to selectional preference modeling, and they prove to be a strong alternative to first-order models. The relative performance of the methods is consistent across the two datasets, stressing the robustness of all methods used.

Regarding the use of similarity software (Padó and Lapata, 2007) on the BNC vs. the use of Lin's ready-made thesaurus, both seem to perform similarly, as exemplified by the similar results of sim_{Lin} and sim^{th}_{Lin}. The fact that the former performed better on the Brown data, and worse on the WSJ data could be related to the different corpora used to compute the co-occurrence, balanced corpus and journalism texts respectively. This could be an indication of the potential of distributional thesauri to adapt to the target domain.

Regarding the similarity metrics, the cosine seems to perform consistently better for first-order distributional similarity, while Jaccard provided slightly better results for second-order similarity.

The best overall performance was for second-order similarity, also using the cosine. Given the computational complexity involved in building a complete thesaurus based on the similarity software, we used the ready-made thesaurus of Lin, but could not try the second-order version on BNC.

6 Conclusions and Future Work

We have empirically shown how automatically generated selectional preferences, using WordNet and distributional similarity measures, are able to effectively generalize lexical features and, thus, improve classification performance in a large-scale argument classification task on the CoNLL-2005 dataset. The experiments show substantial gains on recall and F_1 compared to lexical matching, both on the in-domain WSJ test and, especially, on the out-of-domain Brown test.

Alternative selectional models were studied and compared. WordNet-based models attain good levels of precision but lower recall than distributional similarity methods. A new second-order similarity method proposed in this paper attains the best results overall in all datasets.

The evidence gathered in this paper suggests that using semantic knowledge in the form of selectional preferences has a high potential for improving the results of a full system for SRL, specially when training data is scarce or when applied to out-of-domain corpora.

Current efforts are devoted to study the integration of the selectional preference models presented in this paper in a in-house SRL system. We are particularly interested in domain adaptation, and whether distributional similarities can profit from domain corpora for better performance.

Acknowledgments

This work has been partially funded by the EU Commission (project KYOTO ICT-2007-211423) and Spanish Research Department (project KNOW TIN2006-15049-C03-01). Beñat enjoys a PhD grant from the University of the Basque Country.

References

Carsten Brockmann and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the 10th Conference of the European Chapter of the ACL*, pages 27–34.

X. Carreras and L. Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, MI, USA.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223, Prague, Czech Republic.

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

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, June.

Patrick Pantel and Dekang Lin. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proceedings of the 38th Annual Conference of the ACL*, pages 101–108.

S. Pradhan, W. Ward, and J. H. Martin. 2008. Towards robust semantic role labeling. *Computational Linguistics*, 34(2).

Philip Resnik. 1993. Semantic classes and syntactic ambiguity. In *Proceedings of the workshop on Human Language Technology*, pages 278–283, Morristown, NJ, USA.