Robustness and Generalization of Role Sets: PropBank vs. VerbNet

Beñat Zapirain and **Eneko Agirre**
IXA NLP Group
University of the Basque Country
{benat.zapirain,e.agirre}@ehu.es

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

Abstract

This paper presents an empirical study on the robustness and generalization of two alternative role sets for semantic role labeling: PropBank numbered roles and VerbNet thematic roles. By testing a state–of–the–art SRL system with the two alternative role annotations, we show that the PropBank role set is more robust to the lack of verb–specific semantic information and generalizes better to infrequent and unseen predicates. Keeping in mind that thematic roles are better for application needs, we also tested the best way to generate VerbNet annotation. We conclude that tagging first PropBank roles and mapping into VerbNet roles is as effective as training and tagging directly on VerbNet, and more robust for domain shifts.

1 Introduction

Semantic Role Labeling is the problem of analyzing clause predicates in open text by identifying arguments and tagging them with semantic labels indicating the role they play with respect to the verb. Such sentence–level semantic analysis allows to determine "who" did "what" to "whom", "when" and "where", and, thus, characterize the participants and properties of the *events* established by the predicates. This kind of semantic analysis is very interesting for a broad spectrum of NLP applications (information extraction, summarization, question answering, machine translation, etc.), since it opens the door to exploit the semantic relations among linguistic constituents.

The properties of the semantically annotated corpora available have conditioned the type of research and systems that have been developed so far. PropBank (Palmer et al., 2005) is the most widely used corpus for training SRL systems, probably because it contains running text from the Penn Treebank corpus with annotations on all verbal predicates. Also, a few evaluation exercises on SRL have been conducted on this corpus in the CoNLL-2004 and 2005 conferences. However, a serious criticisms to the PropBank corpus refers to the role set it uses, which consists of a set of numbered core arguments, whose semantic translation is verb-dependent. While Arg0 and Arg1 are intended to indicate the general roles of Agent and Theme, other argument numbers do not generalize across verbs and do not correspond to general semantic roles. This fact might compromise generalization and portability of SRL systems, especially when the training corpus is small.

More recently, a mapping from PropBank numbered arguments into VerbNet thematic roles has been developed and a version of the PropBank corpus with thematic roles has been released (Loper et al., 2007). Thematic roles represent a compact set of verb-independent general roles widely used in linguistic theory (e.g., Agent, Theme, Patient, Recipient, Cause, etc.). We foresee two advantages of using such thematic roles. On the one hand, statistical SRL systems trained from them could generalize better and, therefore, be more robust and portable, as suggested in (Yi et al., 2007). On the other hand, roles in a paradigm like VerbNet would allow for inferences over the assigned roles, which is only possible in a more limited way with PropBank.

In a previous paper (Zapirain et al., 2008), we presented a first comparison between the two previous role sets on the SemEval-2007 Task 17 corpus (Pradhan et al., 2007). The SemEval-2007 corpus only

Proceedings of ACL-08: HLT, pages 550–558,
Columbus, Ohio, USA, June 2008.

comprised examples about 50 different verbs. The results of that paper were, thus, considered preliminary, as they could depend on the small amount of data (both in training data and number of verbs) or the specific set of verbs being used. Now, we extend those experiments to the entire PropBank corpus, and we include two extra experiments on domain shifts (using the Brown corpus as test set) and on grouping VerbNet labels. More concretely, this paper explores two aspects of the problem. First, having in mind the claim that general thematic roles should be more robust to changing domains and unseen predicates, we study the performance of a state-of-the-art SRL system trained on either codification of roles and some specific settings, i.e. including/excluding verb-specific information, labeling unseen verb predicates, or domain shifts. Second, assuming that application scenarios would prefer dealing with general thematic role labels, we explore the best way to label a text with thematic roles, namely, by training directly on VerbNet roles or by using the PropBank SRL system and perform a posterior mapping into thematic roles.

The results confirm our preliminary findings (Zapirain et al., 2008). We observe that the PropBank roles are more robust in all tested experimental conditions, i.e., the performance decrease is more severe for VerbNet. Besides, tagging first PropBank roles and then mapping into VerbNet roles is as effective as training and tagging directly on VerbNet, and more robust for domain shifts.

The rest of the paper is organized as follows: Section 2 contains some background on PropBank and VerbNet role sets. Section 3 presents the experimental setting and the base SRL system used for the role set comparisons. In Section 4 the main comparative experiments on robustness are described. Section 5 is devoted to analyze the posterior mapping of PropBank outputs into VerbNet thematic roles, and includes results on domain–shift experiments using Brown as test set. Finally, Sections 6 and 7 contain a discussion of the results.

2 Corpora and Semantic Role Sets

The PropBank corpus is the result of adding a semantic layer to the syntactic structures of Penn Treebank II (Palmer et al., 2005). Specifically, it provides information about predicate-argument structures to all verbal predicates of the Wall Street Journal section of the treebank. The role set is theory–neutral and consists of a set of numbered core arguments (Arg0, Arg1, ..., Arg5). Each verb has a *frameset* listing its allowed role labels and mapping each numbered role to an English-language description of its semantics.

Different senses for a polysemous verb have different framesets, but the argument labels are semantically consistent in all syntactic alternations of the same verb–sense. For instance in "Kevin broke [the window]$_{Arg1}$" and in "[The door]$_{Arg1}$ broke into a million pieces", for the verb *broke.01*, both Arg1 arguments have the same semantic meaning, that is "broken entity". Nevertheless, argument labels are not necessarily consistent across different verbs (or verb senses). For instance, the same Arg2 label is used to identify the Destination argument of a proposition governed by the verb *send* and the Beneficiary argument of the verb *compose*. This fact might compromise generalization of systems trained on PropBank, which might be focusing too much on verb–specific knowledge. It is worth noting that the two most frequent arguments, Arg0 and Arg1, are intended to indicate the general roles of Agent and Theme and are usually consistent across different verbs. However, this correspondence is not total. According to the study by (Yi et al., 2007), Arg0 corresponds to Agent 85.4% of the time, but also to Experiencer (7.2%), Theme (2.1%), and Cause (1.9%). Similarly, Arg1 corresponds to Theme in 47.0% of the occurrences but also to Topic (23.0%), Patient (10.8%), and Product (2.9%), among others. Contrary to core arguments, adjuncts (Temporal and Location markers, etc.) are annotated with a closed set of general and verb-independent labels.

VerbNet (Kipper et al., 2000) is a computational verb lexicon in which verbs are organized hierarchically into classes depending on their syntactic/semantic linking behavior. The classes are based on Levin's verb classes (Levin, 1993) and each contains a list of member verbs and a correspondence between the shared syntactic frames and the semantic information, such as thematic roles and selectional constraints. There are 23 thematic roles (Agent, Patient, Theme, Experiencer, Source, Beneficiary, Instrument, etc.) which, unlike the Prop-

Bank numbered arguments, are considered as general verb-independent roles.

This level of abstraction makes them, in principle, better suited (compared to PropBank numbered arguments) for being directly exploited by general NLP applications. But, VerbNet by itself is not an appropriate resource to train SRL systems. As opposed to PropBank, the number of tagged examples is far more limited in VerbNet. Fortunately, in the last years a twofold effort has been made in order to generate a large corpus fully annotated with thematic roles. Firstly, the SemLink[1] resource (Loper et al., 2007) established a mapping between PropBank framesets and VerbNet thematic roles. Secondly, the SemLink mapping was applied to a representative portion of the PropBank corpus and manually disambiguated (Loper et al., 2007). The resulting corpus is currently available for the research community and makes possible comparative studies between role sets.

3 Experimental Setting

3.1 Datasets

The data used in this work is the benchmark corpus provided by the SRL shared task of CoNLL-2005 (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. From the input information, we used part of speech tags and full parse trees (generated using Charniak's parser) and discarded named entities. Also, we used the publicly available SemLink mapping from PropBank into VerbNet roles (Loper et al., 2007) to generate a replicate of the CoNLL-2005 corpus containing also the VerbNet annotation of roles.

Unfortunately, SemLink version 1.0 does not cover all propositions and arguments in the PropBank corpus. In order to have an homogeneous corpus and not to bias experimental evaluation, we decided to discard all incomplete examples and keep only those propositions that were 100% mapped into VerbNet roles. The resulting corpus contains 56% of the original propositions, that is, over 50,000 propositions in the training set. This subcorpus is much larger than the SemEval-2007 Task 17 dataset used in our previous experimental work (Zapirain et al., 2008). The difference is especially noticeable in the diversity of predicates represented. In this case, there are 1,709 different verbs (1,505 lemmas) compared to the 50 verbs of the SemEval corpus. We believe that the size and richness of this corpus is enough to test and extract reliable conclusions on the robustness and generalization across verbs of the role sets under study.

In order to study the behavior of both role sets in out–of–domain data, we made use of the PropBanked Brown corpus (Marcus et al., 1994) for testing, as it is also mapped into VerbNet thematic roles in the SemLink resource. Again, we discarded those propositions that were not entirely mapped into thematic roles (45%).

3.2 SRL System

Our basic Semantic Role Labeling system represents the tagging problem as a Maximum Entropy Markov Model (MEMM). The system uses full syntactic information to select a sequence of constituents from the input text and tags these tokens with Begin/Inside/Outside (BIO) labels, using state-of-the-art classifiers and features. The system achieves very good performance in the CoNLL-2005 shared task dataset and in the SRL subtask of the SemEval-2007 English lexical sample task (Zapirain et al., 2007). Check this paper for a complete description of the system.

When searching for the most likely state sequence, the following constraints are observed[2]:

1. No duplicate argument classes for Arg0–Arg5 PropBank (or VerbNet) roles are allowed.
2. If there is a R-X argument (reference), then there has to be a X argument before (referent).
3. If there is a C-X argument (continuation), then there has to be a X argument before.
4. Before a I-X token, there has to be a B-X or I-X token.
5. Given a predicate, only the arguments described in its PropBank (or VerbNet) lexical entry (i.e., the verbal frameset) are allowed.

[1] http://verbs.colorado.edu/semlink/

[2] Note that some of the constraints are dependent of the role set used, i.e., PropBank or VerbNet

Regarding the last constraint, the lexical entries of the verbs were constructed from the training data itself. For instance, the verb *build* appears with four different PropBank core roles (Arg0–3) and five VerbNet roles (Product, Material, Asset, Attribute, Theme), which are the only ones allowed for that verb at test time. Note that in the cases where the verb sense was known we could constraint the possible arguments to those that appear in the lexical entry of that sense, as opposed of using the arguments that appear in all senses.

4 On the Generalization of Role Sets

We first seek a basic reference of the comparative performance of the classifier on each role set. We devised two settings based on our dataset. In the first setting ('SemEval') we use all the available information provided in the corpus, including the verb senses in PropBank and VerbNet. This information was available both in the training and test, and was thus used as an additional feature by the classifier and to constrain further the possible arguments when searching for the most probable Viterbi path. We call this setting 'SemEval' because the SemEval-2007 competition (Pradhan et al., 2007) was performed using this configuration.

Being aware that, in a real scenario, the sense information will not be available, we devised the second setting ('CoNLL'), where the hand-annotated verb sense information was discarded. This is the setting used in the CoNLL 2005 shared task (Carreras and Màrquez, 2005).

The results for the first setting are shown in the 'SemEval setting' rows of Table 1. The correct, excess, missed, precision, recall and F_1 measures are reported, as customary. The significance intervals for F_1 are also reported. They have been obtained with bootstrap resampling (Noreen, 1989). F_1 scores outside of these intervals are assumed to be significantly different from the related F_1 score ($p < 0.05$). The results for PropBank are slightly better, which is reasonable, as the number of labels that the classifier has to learn in the case of VerbNet should make the task harder. In fact, given the small difference, one could think that VerbNet labels, being more numerous, are easier to learn, perhaps because they are more consistent across verbs.

In the second setting ('CoNLL setting' row in the same table) the PropBank classifier degrades slightly, but the difference is not statistically significant. On the contrary, the drop of 1.6 points for VerbNet is significant, and shows greater sensitivity to the absence of the sense information for verbs. One possible reason could be that the VerbNet classifier is more dependant on the argument filter (i.e., the 5th constraint in Section 3.2, which only allows roles that occur in the verbal frameset) used in the Viterbi search, and lacking the sense information makes the filter less useful. In fact, we have attested that the 5th constrain discard more than 60% of the possible candidates for VerbNet, making the task of the classifier easier.

In order to test this hypothesis, we run the CoNLL setting with the 5th constraint disabled (that is, allowing any argument). The results in the 'CoNLL setting (no 5th)' rows of Table 1 show that the drop for PropBank is negligible and not significant, while the drop for VerbNet is more important, and statistically significant.

Another view of the data is obtained if we compute the F_1 scores for core arguments and adjuncts separately (last two columns in Table 1). The performance drop for PropBank in the first three rows is equally distributed on both core arguments and adjuncts. On the contrary, the drop for VerbNet roles is more acute in core arguments (3.7 points), while adjuncts with the 5th constraint disabled get results close to the SemEval setting. These results confirm that the information in the verbal frameset is more important in VerbNet than in PropBank, as only core arguments are constrained in the verbal framesets. The explanation could stem from the fact that current SRL systems rely more on syntactic information than pure semantic knowledge. While PropBank arguments Arg0–5 are easier to distinguish on syntactic grounds alone, it seems quite difficult to distinguish among roles like Theme and Topic unless we have access to the specific verbal frameset. This corresponds nicely with the performance drop for VerbNet when there is less information about the verb in the algorithm (i.e., sense or frameset).

We further analyzed the results by looking at each of the individual core arguments and adjuncts. Table 2 shows these results on the CoNLL setting. The performance for the most frequent roles is similar

PropBank								
Experiment	correct	excess	missed	precision	recall	F_1	F_1 core	F_1 adj.
SemEval setting	6,022	1,378	1,722	81.38	77.76	79.53 $\pm$0.9	82.25	72.48
CoNLL setting	5,977	1,424	1,767	80.76	77.18	78.93 $\pm$0.9	81.64	71.90
CoNLL setting (no 5th)	5,972	1,434	1,772	80.64	77.12	78.84 $\pm$0.9	81.49	71.50
No verbal features	5,557	1,828	2,187	75.25	71.76	73.46 $\pm$1.0	74.87	70.11
Unseen verbs	267	89	106	75.00	71.58	73.25 $\pm$4.0	76.21	64.92

VerbNet								
Experiment	correct	excess	missed	precision	recall	F_1	F_1 core	F_1 adj.
SemEval setting	5,927	1,409	1,817	80.79	76.54	78.61 $\pm$0.9	81.28	71.83
CoNLL setting	5,816	1,548	1,928	78.98	75.10	76.99 $\pm$0.9	79.44	70.20
CoNLL setting (no 5th)	5,746	1,669	1,998	77.49	74.20	75.81 $\pm$0.9	77.60	71.67
No verbal features	4,679	2,724	3,065	63.20	60.42	61.78 $\pm$0.9	59.19	69.95
Unseen verbs	207	136	166	60.35	55.50	57.82 $\pm$4.3	55.04	63.41

Table 1: Basic results using PropBank (top) and VerbNet (bottom) role sets on different settings.

for both. Arg0 gets 88.49, while Agent and Experiencer get 87.31 and 87.76 respectively. Arg2 gets 79.91, but there is more variation on Theme, Topic and Patient (which get 75.46, 85.70 and 78.64 respectively).

Finally, we grouped the results according to the frequency of the verbs in the training data. Table 3 shows that both PropBank and VerbNet get decreasing results for less frequent verbs. PropBank gets better results in all frequency ranges, except for the most frequent, which contains a single verb (*say*).

Overall, the results on this section point out at the weaknesses of the VerbNet role set regarding robustness and generalization. The next sections examine further its behavior.

4.1 Generalization to Unseen Predicates

In principle, the PropBank core roles (Arg0–4) get a different interpretation depending of the verb, that is, the meaning of each of the roles is described separately for each verb in the PropBank framesets. Still, the annotation criteria used with PropBank tried to make the two main roles (Arg0 and Arg1, which account for most of the occurrences) consistent across verbs. On the contrary, in VerbNet all roles are completely independent of the verb, in the sense that the interpretation of the role does not vary across verbs. But, at the same time, each verbal entry lists the possible roles it accepts, and the combinations allowed.

This experiment tests the sensitivity of the two approaches when the SRL system encounters a verb which does not occur in the training data. In principle, we would expect the VerbNet semantic labels, which are more independent across verbs, to be more robust at tagging new predicates. It is worth noting that this is a realistic scenario, even for the verb-specific PropBank labels. Predicates which do not occur in the training data, but do have a PropBank lexicon entry, could appear quite often in the text to be analyzed.

For this experiment, we artificially created a test set for unseen verbs. We chose 50 verbs at random, and split them into 40 verbs for training and 10 for testing (yielding 13,146 occurrences for training and 2,723 occurrences for testing; see Table 4).

The results obtained after training and testing the classifier are shown in the last rows in Table 1. Note that they are not directly comparable to the other results mentioned so far, as the train and test sets are smaller. Figures indicate that the performance of the PropBank argument classifier is considerably higher than the VerbNet classifier, with a $\sim$15 point gap.

This experiment shows that lacking any information about verbal head, the classifier has a hard time to distinguish among VerbNet roles. In order to confirm this, we performed the following experiment.

4.2 Sensitivity to Verb-dependent Features

In this experiment we want to test the sensitivity of the role sets when the classifier does not have any information of the verb predicate. We removed from the training and testing data all the features which make any reference to the verb, including, among others: the surface form, lemma and POS of the verb, and all the combined features that include the verb form (please, refer to (Zapirain et al., 2007) for a complete description of the feature set).

The results are shown in the 'No verbal features'

	CoNLL setting				No verb features	
	PBank		VNet		PBank	VNet
	corr.	F_1	corr.	F_1	F_1	F_1
Overall	5977	78.93	5816	76.99	73.46	61.78
Arg0	1919	88.49			84.02	
Arg1	2240	79.81			73.29	
Arg2	303	65.44			48.58	
Arg3	10	52.63			14.29	
Actor1			44	85.44		0.00
Actor2			10	71.43		25.00
Agent			1603	87.31		77.21
Attribut.			25	71.43		50.79
Cause			51	62.20		5.61
Experien.			215	87.76		86.69
Location			31	64.58		25.00
Patient1			38	67.86		5.71
Patient			208	78.64		25.06
Patient2			21	67.74		43.33
Predicate			83	62.88		28.69
Product			44	61.97		2.44
Recipient			85	79.81		62.73
Source			29	60.42		30.95
Stimulus			39	63.93		13.70
Theme			1021	75.46		52.14
Theme1			20	57.14		4.44
Theme2			21	70.00		23.53
Topic			683	85.70		73.58
ADV	132	53.44	129	52.12	52.67	53.31
CAU	13	53.06	13	52.00	53.06	45.83
DIR	22	53.01	27	56.84	40.00	46.34
DIS	133	77.78	137	79.42	77.25	78.34
LOC	126	61.76	126	61.02	59.56	57.34
MNR	109	58.29	111	54.81	52.99	51.49
MOD	249	96.14	248	95.75	96.12	95.57
NEG	124	98.41	124	98.80	98.41	98.01
PNC	26	44.07	29	44.62	38.33	41.79
TMP	453	75.00	450	73.71	73.06	73.89

Table 2: Detailed results on the CoNLL setting. Reference arguments and verbs have been omitted for brevity, as well as those with less than 10 occ. The last two columns refer to the results on the CoNLL setting with no verb features.

Freq.	PBank	VNet	Freq.	PBank	VNet
0-50	74,21	71,11	500-900	77,97	75,77
50-100	74,79	71,83	> 900	91,83	92,23
100-500	77,16	75,41			

Table 3: F_1 results split according to the frequency of the verb in the training data.

Train	*affect, announce, ask, attempt, avoid, believe, build, care, cause, claim, complain, complete, contribute, describe, disclose, enjoy, estimate, examine, exist, explain, express, feel, fix, grant, hope, join, maintain, negotiate, occur, prepare, promise, propose, purchase, recall, receive, regard, remember, remove, replace, say*
Test	*allow, approve, buy, find, improve, kill, produce, prove, report, rush*

Table 4: Verbs used in the *unseen verb* experiment

rows of Table 1. The performance drops more than 5 points in PropBank, but the drop for VerbNet is dramatic, with more than 15 points.

A closer look at the detailed role-by-role performances can be done if we compare the F_1 rows in the CoNLL setting and in the 'no verb features' setting in Table 2. Those results show that both Arg0 and Arg1 are quite robust to the lack of target verb information, while Arg2 and Arg3 get more affected. Given the relatively low number of Arg2 and Arg3 arguments, their performance drop does not affect so much the overall PropBank performance. In the case of VerbNet, the picture is very different. Focusing on the most frequent roles first, while the performance drop for Experiencer, Agent and Topic is of 1, 10 and 12 points respectively, the other roles get very heavy losses (e.g. Theme and Patient drop 23 and 50 points), and the rest of roles are barely found. It is worth noting that the adjunct labels get very similar performances in both PropBank and VerbNet cases. In fact, Table 1 in the last two rows shows very clearly that the performance drop is caused by the core arguments.

The better robustness of the PropBank roles can be explained by the fact that, when creating PropBank, the human PropBank annotators tried to be consistent when tagging Arg0 and Arg1 across verbs. We also think that both Arg0 and Arg1 can be detected quite well relying on unlexicalized syntactic features only, that is, not knowing which are the verbal and nominal heads. On the other hand, distinguishing between Arg2–4 is more dependant on the subcategorization frame of the verb, and thus more sensitive to the lack of verbal information.

In the case of VerbNet, the more fine-grained distinction among roles seems to depend more on the meaning of the predicate. For instance, distinguishing between Agent–Experiencer, or Theme–Topic–Patient. The lack of the verbal head makes it much more difficult to distinguish among those roles. The same phenomena can be observed among the roles not typically realized as Subject or Object such as Recipient, Source, Product, or Stimulus.

5 Mapping into VerbNet Thematic Roles

As mentioned in the introduction, the interpretation of PropBank roles depends on the verb, and that

Test on WSJ	all	core	adj.
PropBank to VerbNet (hand)	79.17 ±0.9	81.77	72.50
VerbNet (SemEval setting)	78.61 ±0.9	81.28	71.84
PropBank to VerbNet (MF)	77.15 ±0.9	79.09	71.90
VerbNet (CoNLL setting)	76.99 ±0.9	79.44	70.88
Test on Brown			
PropBank to VerbNet (MF)	64.79 ±1.0	68.93	55.94
VerbNet (CoNLL setting)	62.87 ±1.0	67.07	54.69

Table 5: Results on VerbNet roles using two different strategies. Topmost 4 rows for the usual test set (WSJ), and the 2 rows below for the Brown test set.

makes them less suitable for NLP applications. On the other hand, VerbNet roles have a direct interpretation. In this section, we test the performance of two different approaches to tag input sentences with VerbNet roles: (1) train on corpora tagged with VerbNet, and tag the input directly; (2) train on corpora tagged with PropBank, tag the input with PropBank roles, and use a PropBank to VerbNet mapping to output VerbNet roles.

The results for the first approach are already available (cf. Table 1). For the second approach, we just need to map PropBank roles into VerbNet roles using SemLink (Loper et al., 2007). We devised two experiments. In the first one we use the hand-annotated verb class in the test set. For each predicate we translate PropBank roles into VerbNet roles making use of the SemLink mapping information corresponding to that verb lemma and its verbal class.

For instance, consider an occurrence of *allow* in a test sentence. If the occurrence has been manually annotated with the VerbNet class 29.5, we can use the following entry in SemLink to add the VerbNet role Predicate to the argument labeled with Arg1, and Agent to the Arg0 argument.

```
<predicate lemma="allow">
  <argmap pb-roleset="allow.01" vn-class="29.5">
    <role pb-arg="1" vn-theta="Predicate" />
    <role pb-arg="0" vn-theta="Agent" />
  </argmap>
</predicate>
```

The results obtained using the hand-annotated VerbNet classes (and the SemEval setting for PropBank), are shown in the first row of Table 5. If we compare these results to those obtained by VerbNet in the SemEval setting (second row of Table 5), they are 0.5 points better, but the difference is not statistically significant.

experiment	corr.	F_1
Grouped (CoNLL Setting)	5,951	78.11±0.9
PropBank to VerbNet to Grouped	5,970	78.21±0.9

Table 6: Results for VerbNet grouping experiments.

In a second experiment, we discarded the sense annotations from the dataset, and tried to predict the VerbNet class of the target verb using the most frequent class for the verb in the training data. Surprisingly, the accuracy of choosing the most frequent class is 97%. In the case of *allow* the most frequent class is 29.5, so we would use the same SemLink entry as above. The third row in Table 5 shows the results using the most frequent VerbNet class (and the CoNLL setting for PropBank). The performance drop compared to the use of the hand-annotated VerbNet class is of 2 points and statistically significant, and 0.2 points above the results obtained using VerbNet directly on the same conditions (fourth row of the same Table).

The last two rows in table 5 show the results when testing on the the Brown Corpus. In this case, the difference is larger, 1.9 points, and statistically significant in favor of the mapping approach. These results show that VerbNet roles are less robust to domain shifts. The performance drop when moving to an out–of–domain corpus is consistent with previously published results (Carreras and Màrquez, 2005).

5.1 Grouping experiments

VerbNet roles are more numerous than PropBank roles, and that, in itself, could cause a drop in performance. Motivated by the results in (Yi et al., 2007), we grouped the 23 VerbNet roles in 7 coarser role groups. Note that their groupings are focused on the roles which map to PropBank Arg2. In our case we are interested in a more general grouping which covers all VerbNet roles, so we added two additional groups (Agent-Experiencer and Theme-Topic-Patient). We re-tagged the roles in the datasets with those groups, and then trained and tested our SRL system on those grouped labels. The results are shown in the first row of Table 6. In order to judge if our groupings are easier to learn, we can see that he performance gain with respect to the ungrouped roles (fourth row of Table 5) is small (76.99

vs. 78.11) but significant. But if we compare them to the results of the PropBank to VerbNet mapping, where we simply substitute the fine-grained roles by their corresponding groups, we see that they still lag behind (second row in Table 6).

Although one could argue that better motivated groupings could be proposed, these results indicate that the larger number of VerbNet roles does not explain in itself the performance difference when compared to PropBank.

6 Related Work

As far as we know, there are only two other works performing comparisons of alternative role sets on a common test data. Gildea and Jurafsky (2002) mapped FrameNet frame elements into a set of *abstract thematic roles* (i.e., more general roles such as Agent, Theme, Location), and concluded that their system could use these thematic roles without degradation in performance.

(Yi et al., 2007) is a closely related work. They also compare PropBank and VerbNet role sets, but they focus on the performance of Arg2. They show that splitting Arg2 instances into subgroups based on VerbNet thematic roles improves the performance of the PropBank-based classifier. Their claim is that since VerbNet uses argument labels that are more consistent across verbs, they would provide more consistent training instances which would generalize better, especially to new verbs and genres. In fact they get small improvements in PropBank (WSJ) and a large improvement when testing on Brown.

An important remark is that Yi et al. use a combination of grouped VerbNet roles (for Arg2) and PropBank roles (for the rest of arguments). In contrast, our study compares both role sets as they stand, without modifications or mixing. Another difference is that they compare the systems based on the PropBank roles —by mapping the output VerbNet labels back to PropBank Arg2— while in our case we decided to do just the contrary (i.e., mapping PropBank output into VerbNet labels and compare there). As we already said, we think that VerbNet–based labels can be more useful for NLP applications, so our target is to have a SRL system that provides VerbNet annotations. While not in direct contradiction, both studies show different angles of the complex relation between the two role sets.

7 Conclusion and Future work

In this paper we have presented a study of the performance of a state-of-the-art SRL system trained on two alternative codifications of roles (PropBank and VerbNet) and some particular settings, e.g., including/excluding verb–specific information in features, labeling of infrequent and unseen verb predicates, and domain shifts. We observed that PropBank labeling is more robust in all previous experimental conditions, showing less performance drops than VerbNet labels.

Assuming that application-based scenarios would prefer dealing with general thematic role labels, we explore the best way to label a text with VerbNet thematic roles, namely, by training directly on VerbNet roles or by using the PropBank SRL system and performing a posterior mapping into thematic roles. While results are similar and not statistically significant in the WSJ test set, when testing on the Brown out–of–domain test set the difference in favor of PropBank plus mapping step is statistically significant. We also tried to map the fine-grained VerbNet roles into coarser roles, but it did not yield better results than the mapping from PropBank roles. As a side-product, we show that a simple most frequent sense disambiguation strategy for verbs is sufficient to provide excellent results in the PropBank to VerbNet mapping.

Regarding future work, we would like to explore ways to improve the performance on VerbNet roles, perhaps using selectional preferences. We also want to work on the adaptation to new domains of both roles sets.

Acknowledgements

We are grateful to Martha Palmer and Edward Loper for kindly providing us with the SemLink mappings. This work has been partially funded by the Basque Government (IT-397-07) and by the Ministry of Education (KNOW TIN2006-15049, OpenMT TIN2006-15307-C03-02). Beñat is supported by a PhD grant from the University of the Basque Country.

References

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In Ido Dagan and Daniel Gildea, editors, *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, USA, June. Association for Computational Linguistics.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, July.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago.

Edward Loper, Szu-Ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: annotating predicate argument structure. In *HLT '94: Proceedings of the workshop on Human Language Technology*, pages 114–119, Morristown, NJ, USA. Association for Computational Linguistics.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June. Association for Computational Linguistics.

Szu-Ting Yi, Edward Loper, and Martha Palmer. 2007. Can semantic roles generalize across genres? In *Proceedings of the Human Language Technology Conferences/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL-2007)*.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2007. Sequential SRL Using Selectional Preferences. An Approach with Maximum Entropy Markov Models. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 354–357.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2008. A Preliminary Study on the Robustness and Generalization of Role Sets for Semantic Role Labeling. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2008)*, pages 219–230, Haifa, Israel, February.