A Preliminary Study on the Robustness and Generalization of Role Sets for Semantic Role Labeling

Beñat Zapirain[1], Eneko Agirre[1], and Lluís Màrquez[2]

[1] IXA NLP Group
University of The Basque Country
{benat.zapirain,e.agirre}@ehu.es
[2] TALP Research Center
Technical University of Catalonia
lluism@lsi.upc.edu

Abstract. Most Semantic Role Labeling (SRL) systems rely on available annotated corpora, being PropBank the most widely used corpus so far. Propbank role set is based on theory-neutral numbered arguments, which are linked to fine grained verb-dependant semantic roles through the verb framesets. Recently, thematic roles from the computational verb lexicon VerbNet have been suggested to be more adequate for generalization and portability of SRL systems, since they represent a compact set of verb-independent general roles widely used in linguistic theory. Such thematic roles could also put SRL systems closer to application needs. This paper presents a comparative study of the behavior of a state-of-the-art SRL system on both role role sets based on the SemEval-2007 English dataset, which comprises the 50 most frequent verbs in PropBank.

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 avenue for exploiting the semantic relations among linguistic constituents.

The increasing availability of large semantically annotated corpora, like PropBank and FrameNet, has contributed to increase the interest on the automatic development of Semantic Role Labeling systems in the last five years. Since Gildea and Jurafsky's initial work "Automatic Labeling of Semantic Roles" [3] on FrameNet-based SRL, many researchers have devoted their efforts on this exciting and relatively new task. Two evaluation exercises on SRL were conducted by the 'shared tasks' of CoNLL-2004 and CoNLL-2005 conferences [1, 2], bringing to scene a comparative analysis of almost 30 competitive systems trained on the PropBank corpus. From there, PropBank became the most widely used corpus for training SRL systems.

One of the 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, specially when the training corpus is small and not very representative. Thematic roles (e.g., based on VerbNet) have been suggested to be more adequate for generalization and portability, since they represent a compact set of verb-independent general roles widely used in linguistic theory. Such thematic roles could also put SRL systems closer to application needs [11].

Thanks to a mapping from PropBank numbered arguments into VerbNet thematic roles, a version of the PropBank corpus with thematic roles has been released recently [6]. Using a part of this corpus, an English SRL task was proposed in SemEval-2007, which compared the results of the systems under both role sets [9]. Unfortunately, the number of participants in that task was too small to extract reliable conclusions.

In this paper, we go further in this direction and describe an experimental comparison between the two previous role sets (PropBank numbered arguments vs. VerbNet thematic roles). 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 training on either codification of roles and some specific settings, e.g., including/excluding verb-specific information in features, and labeling of unseen verb predicates. Although numerical results are not directly comparable we observe that the PropBank-based labeling is more robust in all previous experimental conditions (i.e., the performance decrease is less severe than in the VerbNet case). Finally, assuming that application-based 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 rest of the paper is organized as follows: Section 2 contains background on PropBank and VerbNet-based thematic roles. Section 3 presents the experimental setting of our experiments 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-style output into VerbNet thematic roles. Finally, Sections 6 and 7, contain a discussion of the results in context and outline the main directions for future research.

2 Corpora and Semantic Role Sets

The PropBank corpus is the result of adding a shallow semantic layer to the syntactic structures of Penn Treebank II [8]. 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 allowing role labels and mapping each numbered role to an English-language description of the semantics of the role, which is specific to that verb.

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 biased to acquire too verb-specific knowledge. In spite of that fact, and thanks to some annotation criteria, the most frequent arguments in PropBank, Arg0 and Arg1, are intended to indicate the general roles of Agent and Theme and are usually consistent across different verbs. Adjuncts (Temporal and Location markers, etc.) conform also a set of general and verb-independent labels. PropBank has become the most widely used corpus for training SRL systems due to two main reasons: first, PropBank provides a representative sample of general text with complete role-annotations; and second, the numerous international evaluations using PropBank highly promoted its usage among the researchers.

VerbNet [4] 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 [5] and contain semantic and syntactic information about 4,526 verb senses (corresponding to 3,769 lexemes). Each class comprises a list of member verbs and associates their shared syntactic frames with 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 PropBank numbered arguments, are considered as general verb-independent roles.

This level of abstraction makes them, in principle, more suited than PropBank numbered arguments for being directly exploited by general NLP applications. But, VerbNet by itself is not an appropriate lexical 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[3] resource [6] 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 [6]. The resulting corpus is currently available for the research community and makes possible comparative studies between role sets like [11] and the one in this paper.

3 Experimental Setting

3.1 Datasets

The data used in the experiments is the one provided by the SRL subtask of the English lexical sample in SemEval-2007[4]. The dataset comprises the occurrences of 50 different

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

[4] http://www.cs.swarthmore.edu/semeval

verb lemmas from the WSJ portion of PropBank. It includes the part of speech and full syntactic information for each word as well as the hand tagged PropBank frame sense and the VerbNet class for verbs. The training data is a subsection from Sections 02-21 and the test data comprises Sections 01, 22, 23 and 24.

The corpus is annotated with two different semantic role sets, the PropBank role set and the VerbNet thematic role set. There is a total of 5 (core) role types for PropBank and 213 thematic roles for VerbNet. In a small number of cases, there is no VerbNet role available (e.g. when VerbNet does not contain the appropriate sense of the verb) so the PropBank role label is given instead. Apart from the argument role labels, both versions of the dataset are annotated with common adjunct like roles such as temporal, adverbial, location and so on.

The 50 verbs from the dataset cover a wide range of VerbNet classes (see table 1). Therefore, most of the classes are not strongly represented in the training set because of the relatively small size of the dataset and the large number of covered classes. Table 1 also shows the number of verb occurrences in those classes.

All in all, the training part has an average of 317.36 occurrences per verb, ranging from 8,365 for *say* to 23 for *regard*. The test has an average of 61.88 occurrences per verb, ranging from 1,665 for *say* to 4 for *grant*. The average polisemy for VerbNet is 1.71 and for PropBank is 1.70. The verbs are linked to a total of 44 VerbNet classes, with an average of 1.13 verbs per class.

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 [10]. The system achieves competitive performance in the CoNLL-2005 shared task dataset and ranked first in the SRL subtask of the SemEval-2007 English lexical sample task [12].

Maximum Entropy Markov Models are discriminative models for sequential tagging (i.e., the problem of assigning a sequence of labels $[s_1, \ldots, s_n]$ to a sequence of observations $[o_1, \ldots, o_n]$) that model the local probability distribution $P(s_i \mid s_{i-1}, \hat{o}_i)$ for each possible label s_i at position i, where $\hat{o}_i$ is the context of observation o_i and s_{i-1} the preceding label. Given a MEMM, the most likely state sequence is the one that maximizes the following formula

$$S = \text{argmax}_{[s_1,\ldots,s_n]} \prod_{i=1}^{n} P(s_i \mid s_{i-1}, \hat{o}_i)$$

Translating the problem to SRL, we have role/argument labels connected to each state in the sequence (or proposition), and the observations are the features extracted in these points (token features). We get the most likely label sequence finding out the most likely state sequence (using the Viterbi algorithm). All the conditional probabilities are given by the Maximum Entropy classifier with a tunable Gaussian prior from the Mallet Toolkit[5], which was empirically set to 0.1 in these experiments.

[5] http://mallet.cs.umass.edu

Verb	VN	PB	train	test	Verb	VN	PB	train	test
affect	31.1	01	121	28	feel	30.4	01	6	2
affect	None	02	1	0	feel	30.4	05	1	0
allow	29.5	01	254	42	feel	31.3	03	13	1
allow	29.5-1	03	1	0	find	13.5.1	01	195	29
allow	64	02	4	0	find	29.4	01	170	27
allow	None	02	4	0	find	None	01	9	1
announce	37.7	01	291	41	fix	26.3-1	01	1	
approve	31.3	01	173	35	fix	26.3-1	02	8	
ask	37.1-1	01	118	13	fix	54.4	03	47	8
ask	37.1-1	02	149	20	grant	13.3	01	30	4
ask	None	03	3	1	hope	32.2	01	162	46
attempt	61	01	57	14	improve	45.4	01	149	28
avoid	52	01	102	18	improve	45.4	02	13	
believe	29.4	01	325	54	join	22.1-2	01	120	20
build	26.1-1	01	224	38	kill	42.1-1	01	80	9
build	26.2	02	39	5	maintain	29.5	01	122	13
build	None	03	7	4	negotiate	36.1	01	82	16
build	None	05	5	0	occur	48.3	01	65	21
buy	13.5.1	01	743	16	prepare	26.3-1	01	35	5
buy	13.5.1	02	4	0	prepare	26.3-1	02	53	16
buy	13.5.1	03	3	2	produce	26.4	01	262	52
care	31.3	01	21	4	promise	13.3	01	69	10
cause	27	01	195	46	propose	37.7	01	198	42
claim	37.7	01	106	23	prove	29.4	01	88	21
claim	None	02	2	0	purchase	13.5.2-1	01	135	32
complain	37.8	01	75	13	recall	10.2	01	6	2
complete	55.2	01	167	30	recall	29.2	02	53	6
contribute	13.2	01	103	30	receive	13.5.2	01	326	67
describe	29.2	01	68	11	regard	29.2	01	22	5
disclose	37.7	01	163	28	regard	None	01	1	0
disclose	None	01	4	0	remember	29.2	01	38	5
enjoy	31.2	01	40	8	remove	10.1	01	44	3
estimate	54.4	01	255	45	remove	10.2	01	16	7
examine	35.4	01	20	6	replace	13.6	01	84	25
exist	47.1-1	01	105	11	report	29.1	01	455	99
explain	37.1	01	89	16	report	37.7	01	74	19
express	11.1-1	02	1	0	report	None	01	9	1
express	48.1.2	01	51	11	rush	51.3.2	01	33	7
feel	29.5	02	52	17	say	37.7	01	8,365	1,645
feel	30.1	01	85	19					

Table 1. Verbs in the dataset in alphabetic order. VerbNet (VN) class and Propbank (PB) senses are given, as well as the occurrences in the train and test sets. ‘None’ means that no VerbNet class was assigned

The full list of features used can be found in [12]. From that setting, we excluded the experimental semantic features based on selectional preferences, which could interfere with the interpretation of the results. The features are the same for both PropBank and VerbNet. In both cases a single MEMM classifier is trained for all verbs using all the available training data.

When searching for the most likely state sequence, the following constraints are observed[6]:

1. No duplicate argument classes for Arg0–Arg5 Propbank roles (or VerbNet roles) are allowed.

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

PropBank						
Experiment	correct	excess	missed	precision	recall	F_1
SemEval setting	5,703	1,009	1,228	84.97	82.28	83.60 ±0.9
CoNLL setting	5,690	1,012	1,241	84.90	82.09	83.47 ±0.8
CoNLL setting (no 5th)	5,687	1,019	1,244	84.80	82.05	83.41 ±0.8
No verbal features	5,575	1,134	1,356	83.10	80.44	81.74 ±0.9
Unseen verbs	5,125	1,282	1,639	79.99	75.77	77.82 ±0.9

VerbNet						
Experiment	correct	excess	missed	precision	recall	F_1
SemEval setting	5,681	993	1,250	85.12	81.97	83.51 ±0.9
CoNLL setting	5,650	1,042	1,281	84.43	81.52	82.95 ±0.8
CoNLL setting (no 5th)	5,616	1,106	1,315	83.55	81.03	82.27 ±1.0
No verbal features	4,941	1,746	1,990	73.89	71.29	72.57 ±1.0
Unseen verbs	3,691	2,555	3,073	59.09	54.57	56.74 ±0.9

Table 2. Basic results using PropBank and VerbNet role sets

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 are allowed.

Regarding the last constraint, the lexical entries of the verbs were constructed from the training data itself. For instance, for the verb *build* the PropBank entry would only allow 4 core roles (Arg0-3), while the VerbNet entry would allow 6 roles (Product, Material, Asset, Attribute, Theme and Arg2). Note that in the cases where the PropBank (or VerbNet) sense is known (see below) we would constraint the possible arguments only 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

First, we wanted to have a basic reference of the comparative performance of the classifier on each role set. We performed two experiments. In the first one we use all the available information provided by the SemEval organizers, including the verb senses in PropBank and VerbNet. This information was available both in the test and train data, and was thus used as an additional feature by the classifier and to constraint further the possible arguments when searching for the most probable Viterbi path.

The results are shown in the 'SemEval setting' rows of Table 2. The correct, excess, missed, precision, recall and F_1 measures are reported, as customary. The significance intervals for F_1 are also reported, which have been obtained with bootstrap resampling [7]. F_1 scores outside of these intervals are assumed to be significantly different from the related F_1 score ($p < 0.05$). The precision is higher for VerbNet, but

	SemEval setting								No verb feature			
	PropBank				VerbNet				PropBank		VerbNet	
	corr.	prec.	rec.	F_1	corr.	prec.	rec.	F_1	corr.	F_1	corr.	F_1
Overall	5703	84.97	82.28	83.60	5681	85.12	81.97	83.51	5575	81.74	4941	72.57
Arg0	2507	93.41	92.34	92.87					2492	91.82		
Arg1	2470	83.45	82.64	83.04					2417	81.34		
Arg2	115	72.33	65.71	68.86	0	0.00	0.00	0.00	76	48.10	0	0.00
Arg3	25	60.98	50.00	54.95	8	57.14	47.06	51.61	18	39.56	3	28.57
Arg4	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0	0.00
Actor1					10	90.91	83.33	86.96			0	0.00
Actor2					1	100.00	100.00	100.0			1	66.67
Agent					2357	93.49	92.40	92.94			2339	89.24
Asset					15	68.18	71.43	69.77			12	52.17
Attribute					8	72.73	47.06	57.14			6	46.15
Beneficiary					14	66.67	58.33	62.22			6	33.33
Cause					36	78.26	75.00	76.60			1	3.64
Experiencer					118	90.08	89.39	89.73			5	7.14
Location					9	100.00	75.00	85.71			0	0.00
Material					1	100.00	14.29	25.00			0	0.00
Patient					28	96.55	75.68	84.85			3	14.29
Patient1					17	85.00	85.00	85.00			3	26.09
Predicate					124	73.81	68.51	71.06			58	37.42
Product					73	70.87	68.87	69.86			10	14.49
Recipient					39	88.64	81.25	84.78			36	67.29
Source					15	62.50	60.00	61.22			15	57.69
Stimulus					11	61.11	52.38	56.41			9	45.00
Theme					525	83.20	80.77	81.97			352	47.70
Theme1					52	85.25	75.36	80.00			4	10.39
Theme2					39	72.22	65.00	68.42			1	3.12
Topic					1594	86.16	85.38	85.77			1511	79.30
ArgM-ADV	97	56.40	51.60	53.89	96	55.81	51.06	53.33	97	54.19	95	53.67
ArgM-CAU	2	100.00	15.38	26.67	4	100.00	30.77	47.06	4	44.44	3	35.29
ArgM-DIR	2	100.00	50.00	66.67	2	100.00	50.00	66.67	2	66.67	2	66.67
ArgM-EXT	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0	0.00
ArgM-LOC	104	63.03	68.87	65.82	105	61.40	69.54	65.22	103	64.38	105	65.83
ArgM-MNR	37	49.33	43.53	46.25	38	50.00	44.71	47.20	31	41.89	32	40.25
ArgM-PNC	7	58.33	25.00	35.00	8	57.14	28.57	38.10	5	26.32	6	29.27
ArgM-PRD	0	0.00	0.00	0.00	3	100.00	33.33	50.00	0	0.00	0	0.00
ArgM-REC	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0	0.00
ArgM-TMP	265	74.65	68.65	71.52	263	73.06	68.13	70.51	263	70.79	265	70.48

Table 3. Detailed results on the SemEval setting for PropBank and VerbNet roles, left and right respectively. Excess and missed numbers, as well as reference arguments and verbs have been omitted for brevity. The rightmost rows show the figures for the 'no verb features' setting.

the recall is lower and the F_1 score is slightly better for PropBank. The differences are nevertheless very small, and given the confidence interval for F_1, negligible. The number of labels that the classifier has to learn in the case of VerbNet should make the task harder. Given the fact that the same results are obtained with respect to PropBank could lead one to think that the VerbNet labels are easier to learn, perhaps because they are more consistent across verbs.

In fact, the detailed results for the roles in each of the sets in Table 3 (rows for the SemEval setting) seem to support this fact. The table shows the larger number of roles that need to be learned for VerbNet.

The performance for the most frequent roles is very similar in both sets. For instance, Arg0 and Agent (2,507 and 2,357 correct labels respectively) both have an F_1

score of 92%. Arg1 (with 2,470 correct labels) get 83% of F_1, but VerbNet Topic and Theme (with 1,594 and 525 correct labels) get 85% and 82% of F_1.

In the second experiment we restricted the use of hand annotated information. This setting is more natural, as it does not use any gold standard data in the test part in order to predict the roles. The results are shown in the ‘CoNLL setting’ rows of Table 2. We can see that while the PropBank classifier did not suffer any appreciable loss, the thematic role classifier showed greater sensitivity to the absence of this kind of information. One possible reason could be that the VerbNet classifier is more sensitive to the argument filter (the 5th constraint) used in the Viterbi search, and lacking the sense information makes the filter less useful. In any case, neither differences are significant according to the confidence intervals.

In order to test how important is the 5th constraint, 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 2 show that the drop for PropBank is negligible, but the drop in VerbNet is more important. In fact, the difference in performance from the SemEval setting to that obtained without the VerbNet class and argument constraints is statistically significant.

In the next subsections we examine the robustness and generalization capabilities for each of the role sets.

4.1 Generalization to Unseen Predicates

In principle, the PropBank core roles (Arg0–4) get a different interpretation depending of the verb, i.e. the meaning of each of the roles is described separately for each verb in the PropBank framesets. Still, the annotation criteria set for PropBank tried to make the two main roles accounting for most of the occurrences consistent across verbs. In VerbNet, to the contrary, 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 which combinations are allowed.

This experiment tests the sensitivity of the role sets when the classifier encounters a verb which does not occur in the training data. This is a realistic case, as in many cases, verbs without training data are found in the target corpora to be processes. In principle, we would expect the set which is more independent across verbs to be more robust. We artificially created a test set for unseen verbs. We first chose 10 verbs at random, and removed their occurrences from the training data, yielding 13,146 occurrences for the 40 verbs. In order to have a sizeable test set, we tested on the 2,723 occurrences of those 10 verbs in the train set (see Table 4).

The results obtained after training and testing the classifier are shown in the last rows in Table 2. Note that they are not directly comparable to the other results mentioned so far, as the test set is a subset of the original test set. The figures indicate that the performance of the PropBank argument classifier is considerably higher than the VerbNet classifier, with a 20 point gap.

This experiment shows that not knowing the verbal head, the classifier has a very hard time to distinguish among the fine-grained VerbNet roles. In order to confirm this, we performed further analysis, as described in the next subsection.

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

4.2 Sensitivity to Verb-dependent Features

In this experiment we want to test the sensitivity of the sets when the classifier does not have any information of the main verb in the sentence where it is tagging the argument and adjuncts. 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 [12] for a complete description of the feature set used).

The results are shown in the 'No verbal features' rows of Table 2. The performance drop in PropBank is small, on the fringe of being statistically significant, but the drop for VerbNet is dramatic, 10 points in precision, recall and F_1 with clear statistical significance. A closer look at the detailed role-by-role performances can be done if we compare the F_1 rows in the SemEval setting and in the 'no verb features' setting in Table 3. Those results show that both Arg0 and Arg1 are very 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 much the overall PropBank performance. In the case of VerbNet, the picture is very different. While the performance drop for Agent and Topic is of 2 and 5 points respectively, the other roles get very heavy losses: Theme and Predicate get their F_1 halfed, and the rest of roles are barely found. It is worth noting that the adjunct labels get very similar performances in all cases.

The robustness of the PropBank roles can be explained by the fact that the PropBank taggers 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, i.e. 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 Theme and Recipient, not to say about Theme, Theme1 and Theme2. The lack of the verbal head makes it much more difficult to distinguish among those roles.

5 Mapping into VerbNet Thematic Roles

As mentioned in the introduction, the interpretation of PropBank roles depends on the verb, and that makes them less suitable for NLP applications. VerbNet roles, on the other hand, 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 approximation are already available (cf. Table 2). For the second approximation, we just need to map PropBank roles into VerbNet roles using Semlink [6]. We devised two experiments. In the first one we use the hand-annotated verb class in the test set. For each verb governing a proposition 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 only 0.1 lower, and the difference is not statistically significant.

experiment	corr.	excess	missed	prec.	rec.	F_1
PropBank to VerbNet (hand)	5,680	1,009	1,251	84.92	81.95	83.41 ±0.9
VerbNet (SemEval setting)	5,681	993	1,250	85.12	81.97	83.51 ±0.9
PropBank to VerbNet (most frequent)	5,628	1,074	1,303	83.97	81.20	82.56 ±0.8
VerbNet (CoNLL setting)	5,650	1,042	1,281	84.43	81.52	82.95 ±0.8

Table 5. Results on VerbNet roles using two different strategies. The 'PropBank to VerbNet' rows show the results using the mapping. The results for directly using VerbNet are taken from Table 2.

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. The accuracy of choosing the most frequent class is of 97% on the training. In the case of *allow* the most frequent class is 29.5 (cf. Table 1), 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 small, and barely statistically significant, and only 0.4 from the results obtained directly using VerbNet on the same conditions (fourth row of the same Table).

All in all, the second experiment shows that, in realistic conditions, using VerbNet directly provides the same results than tagging with PropBank roles, disambiguating with the most frequent VerbNet class and then using Semlink for mapping. These results may imply that the classifier is not able to learn better from VerbNet roles rather than PropBank roles.

6 Related Work

As far as we know, there are only two other works doing an extensive comparison of different role sets on the same test data.

Gildea and Jurafsky [3] 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 and Loper [11] is a closely related work, and as far as we know, the only other work doing an extensive comparison of different role sets on the same test data. The authors also compare PropBank and VerbNet role sets, but they focus on the performance of Arg2. The authors show that splitting Arg2 instances into subgroups based on thematic roles improves the performance of the PropBank-based classifier, especially in out-of-domain experiments (Brown corpus).

Note that the authors do not use purely VerbNet roles, but 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, and our results show that VerbNet roles are less robust and not easier to learn than PropBank roles. While not in direct contradiction, both studies show different angles of the complex relation between the different role sets.

7 Conclusion and Future work

In this paper we present a preliminary study of the performance of a state-of-the-art SRL system training on either codification of roles and some specific settings, e.g., including/excluding verb-specific information in features, and labeling of infrequent and unseen verb predicates. We observed that the PropBank-based labeling is more robust in all previous experimental conditions (i.e., the performance decrease is less severe than in the VerbNet case). Finally, assuming that application-based 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. In this case, we find that the difference is not statistically significant.

Regarding future work, we want to extend this work to all the verbs in VerbNet. Among other things, we would like to test whether having more verbs to train affects the relative performance of PropBank and VerbNet. We would also like to improve the results for the VerbNet role set using role groupings in order to reduce the sparsity of the data. Finally, we would like to revisit the portability results of [11] using our setting.

Acknowledgements

This work has been partially funded by the Department of Education, Universities and Research of the Basque Government (IT-397-07) and by the Ministry of Education Projects KNOW and OpenMT (TIN2006-15049-C03-01; TIN2006-15307-C03-02). Beñat is supported by a PhD grant from the University of the Basque Country.

References

1. X. Carreras and L. Màrquez. Introduction to the conll-2004 shared task: Semantic role labeling. In H. Ng and E. Riloff, editors, *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, MA, USA, May 2004. Association for Computational Linguistics.
2. X. Carreras and L. Màrquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In I. Dagan and D. Gildea, editors, *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, USA, June 2005. Association for Computational Linguistics.
3. D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.
4. K. Kipper, H. T. Dang, and M. Palmer. Class based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, July 2000.
5. B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago, 1993.
6. E. Loper, S.-T. Yi, and M. Palmer. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands, 2007.
7. E. W. Noreen. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons, 1989.
8. M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105, 2005.
9. S. Pradhan, E. Loper, D. Dligach, and M. Palmer. 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 2007. Association for Computational Linguistics.
10. M. Surdeanu, L. Màrquez, X. Carreras, and P. R. Comas. Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research (JAIR)*, 29:105–151, 2007.
11. S.-T. Yi, E. Loper, and M. Palmer. 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)*, 2007.
12. B. Zapirain, E. Agirre, and L. Màrquez. Sequential SRL using selectional preferences: An aproach with Maximum Entropy Markov Models. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic, Association for Computational Linguistics*, 2007.