Robust Semantic Analysis for Unseen Data in FrameNet

Alexis Palmer, Afra Alishahi, Caroline Sporleder
Computational Linguistics and Phonetics
Saarland University, Germany
{apalmer,afra,csporled}@coli.uni-saarland.de

Abstract

We present a novel method for FrameNet-based semantic role labeling (SRL), focusing on limitations posed by the limited coverage of available annotated data. Our SRL model is based on Bayesian clustering and has the advantage of being very robust in the face of unseen and incomplete data. Frame labeling and role labeling are modeled in like fashions, allowing cascading classification scenarios. The model is shown to perform especially well on unseen data. In addition, we show that for seen data, predicting semantic types for roles improves role labeling performance.

1 Introduction

The majority of recent work in semantic role labeling (SRL) has been carried out on PropBank-style semantic argument annotations (Palmer et al., 2005), rather than on FrameNet-style annotations (Ruppenhofer et al., 2006). FrameNet differs from PropBank in that FrameNet annotations are more strongly semantically driven. FrameNet generalizes over different parts of speech and can assign the same sense (*frame*) to a noun and a verb as in (1), where both *competition* and *play* are assigned the COMPETITION frame. Also, FrameNet assigns semantic roles not only to syntactic arguments of the target but also to constituents which are not directly syntactically dependent on the target but can be semantically understood as filling a role, e.g., *Wivenhoe Town* in (1a).

(1) a. [Wivenhoe Town]$_{Participant1}$ have never won the **competition**$_{Competition}$.
b. [Olympiakos]$_{Participant1}$ **plays**$_{Competition}$ [against Aris Salonica]$_{Participant1}$ [in Piraeus]$_{Place}$.

A major challenge for FrameNet-style SRL is posed by the limited coverage of available annotated data. The FrameNet lexicographic corpus was annotated on a frame-by-frame basis, selecting individual example sentences for each *lexical unit* (LU), or pairing of lemma and frame. This means that many common lemmas are missing from FrameNet, and for those that *are* included the number of example sentences is often relatively small and not in accordance with distributions found in naturally-occurring texts.

FrameNet's well-known coverage gaps translate directly to drops in labeling performance, motivating the development of systems which are more robust in the face of sparse data. For example, the supervised SRL system Shalmaneser (Erk and Padó, 2006) obtains a frame labeling accuracy of 93% on FrameNet 1.2 (with a 90-10 training-test split), but the same system's performance drops to 47% accuracy when trained on FrameNet 1.3 and tested on texts with full frame-semantic annotations (Palmer and Sporleder, 2010). Similarly, Das et al. (2010) report a 60% frame labeling F-Score on SemEval-07 data, but of 210 unseen lemmas, their system predicts just four frames correctly.[1]

In general the term *unseen* could refer to unseen frames, unseen lemmas, or unseen LUs. As further discussed in Section 4, we are interested in unseen LUs: cases in which the system has not been exposed to a particular pairing of lemma and frame. We propose a novel method for SRL based on Bayesian clustering. The model is well suited to deal with incomplete data, both in terms of missing feature values and in terms of feature-label combinations not seen in the training data.

2 Related Work

While early FrameNet-style SRL systems (Gildea and Jurafsky, 2002; Erk and Padó, 2006, among others) are unable to make predictions for LUs not seen in the training data, several more recent stud-

[1] Under the SemEval-07 partial matching scheme, a majority of the other frame predictions receive partial credit.

Proceedings of Recent Advances in Natural Language Processing, pages 628–633,
Hissar, Bulgaria, 12-14 September 2011.

ies have addressed the coverage issue. For example, Das et al. (2010) introduce a latent variable ranging over seen targets, allowing them to infer likely frames for unseen words, and the SRL system of Johansson and Nugues (2007) uses WordNet to generalise to unseen lemmas. In a similar vein, Burchardt et al. (2005) propose a system that generalizes over WordNet synsets to guess frames for unknown words. Pennacchiotti et al. (2008) compare WordNet-based and distributional approaches to inferring frames and conclude that a combination of the two leads to the best results, while (Cao et al., 2008) discuss how different distributional models can be utilised. Several approaches have also addressed other coverage problems, e.g., how to automatically expand the number of example sentences for a given lexical unit (Padó et al., 2008; Fürstenau and Lapata, 2009).

Another related approach is that of generalizing over semantic roles. Baldewein et al. (2004) use the FrameNet hierarchy to model the similarity of roles, boosting seldom-seen instances by reusing training data for similar roles, though without significant gains in performance. The most extensive study on role generalization to date (Matsubayashi et al., 2009) compares different ways of grouping roles—exploiting hierarchical relations in FrameNet, generalizing via role names, utilising role types, and using thematic roles from VerbNet—with the best results from using all groups together.

3 Model

We formalize frame and role assignment using an extended version of the construction learning model of Alishahi and Stevenson (2010). The model uses Bayesian clustering for learning argument structure constructions: each construction is a grouping of individual predicate usages which probabilistically share form-meaning associations. These groupings typically correspond to general constructions in the language such as intransitive, transitive, and ditransitive. By detecting similar usages and clustering them into constructions, the model forms probabilistic associations between syntactic positions of arguments with respect to the predicate, and the lexical semantic properties of the predicate and the arguments.

We model frame and role assignment in this fashion, where the most probable values for a missing frame or the semantic roles of arguments are predicted based on the acquired constructions (or clusters), and the extracted features from the corpus. This strategy provides a number of advantages. First, the model can easily deal with incomplete data; that is, input instances for which any number of features are missing can be seamlessly clustered or considered for prediction, based on the similarity of their features with those in the existing clusters. Moreover, a single core prediction mechanism is used for a variety of tasks (e.g. predicting a missing frame label, role, or role type), which can lead to cascading prediction. For example, for a partial (i.e. unannotated) frame instance, the best role type for each argument can be predicted based on the available features, and then argument roles can be predicted based on those features and the predicted role types.

An important characteristic of this model is its generalizability. It uses a full Bayesian prediction model, which takes into account the contribution of *every* cluster to predicting the best value for a missing feature. This way, there is no built-in difference between predicting a frame label or semantic role for *seen* versus *unseen* instances. Naturally, the outcome of prediction will be more accurate if the model has seen several instances similar to a test instance (i.e., from the same lexical unit or lemma). But even for unseen instances, the model is still capable of generalizing the properties of the training instances given that there are similarities between their available features, such as the syntactic pattern and the semantic properties of the predicate and the arguments.

3.1 Clustering Frame Instances

From the FrameNet corpus, we extract for each instance the nine features shown in Table 1. Different subsets of these features are used for the experiments reported in Section 5.

An incremental Bayesian clustering process groups each extracted frame instance with the most similar existing cluster of instances. If no existing cluster has sufficiently high probability for the new frame instance, a new cluster is created.

Adding a frame instance X to a cluster c is formulated as finding the c with the maximum probability given X, where c ranges over the indices of all clusters, with index 0 representing recognition of a new cluster. Using Bayes rule, and dropping $P(X)$ which is constant for all c:

$$P(c|X) = \frac{P(c)P(X|c)}{P(X)} \sim P(c)P(X|c) \quad (2)$$

The prior probability $P(c)$ is given by the relative frequency of the frame instances it contains, over all observed instances. The posterior probability of an instance X is expressed in terms of the individual probabilities of its features, which we assume are independent, thus yielding a simple product of feature probabilities:

$$P(X|c) = \prod_{i \in \mathit{Features}(X)} P(X_i|c) \tag{3}$$

This probability is estimated using smoothed maximum likelihood:

$$P(X_i|c) = \frac{\sum_{X' \in c} \mathrm{match}(X_i, X'_i) + \lambda}{n_c + \alpha_i \lambda} \tag{4}$$

where n_c is the number of instances in cluster c, and α_i and λ are the smoothing factors. For single-valued features (e.g. head word), the function match returns 1 if the two feature values are identical, and 0 otherwise.

For features whose value is a set (semantic properties of the predicate and arguments, word classes), an exact match between two sets is rare. We instead assume that the members of set-valued features are independent of each other, and calculate the probability of displaying a set S_i on feature i in cluster c as:

$$P(S_i|c) = \frac{1}{|S_c \cup S_i|}(\prod_{s \in S_i} P(s|c) \times \prod_{s \in S_c - S_i} P(\neg s|c)) \tag{5}$$

where S_c is the superset of all the set values of feature i for members in cluster c. Likelihood probabilities $P(s|c)$ and $P(\neg s|c)$ are estimated as in Eqn. (4), by counting members of cluster c whose value for feature f does or does not contains s, respectively. The product is rescaled by the size of the union of the two sets, $S_c \cup S_i$.

3.2 Frame Identification and Role Assignment

For any instance in the test set, both frame identification and role assignment can be modeled as finding the most probable value for a target feature, given other available features.

The probability of an unobserved feature i displaying value X_i given other feature values in an instance X is estimated as:

$$\begin{aligned} P(X_i|X) &= \sum_c P(X_i|c)P(c|X) \qquad (6) \\ &= \sum_c P(X_i|c)P(c)P(X|c) \end{aligned}$$

The conditional probabilities $P(X|c)$ and $P(X_i|c)$ are determined as in the learning module. Ranging over the possible values X_i of feature i, the value of an unobserved feature can be predicted by maximizing $P(X_i|X)$:

$$\mathrm{BestValue}(X, i) = \operatorname*{argmax}_{X_i} P(X_i|X) \tag{7}$$

Identifying a frame can be simulated as finding the frame label X_{frame} with the highest $P(X_{\mathit{frame}}|X)$, or estimating $\mathrm{BestValue}(X, \mathrm{frame})$. Similarly, assigning roles or role types to the arguments of an instance X is modeled as estimating $\mathrm{BestValue}(X, \mathrm{role})$ or $\mathrm{BestValue}(X, \mathrm{role\_type})$, respectively.

4 Data

In this work we use the FrameNet 1.3 lexicographic corpus to evaluate the performance of our model on both seen and unseen data. This corpus provides annotated example sentences for each lexical unit (LU; frame-lemma pairing), documenting a range of syntactic and semantic usages, and it consists of 139,439 annotated example sentences distributed over 10,195 LUs. After excluding 4161 sentences due to inconsistencies with FrameNet definitions, we created two data sets: **seen** and **unseen**.

Seen Data. In the seen set-up, we assume that the model has complete information about each instance's lexicographic status. This means that for *frame labeling* the model knows which frames each target lemma can have and, further, has access to the training instances for each of those frames. Frame labeling is thus performed on a lemma-by-lemma basis. For *role labeling* we assume that the frame of the target lemma is known (e.g., has been previously predicted, either automatically or by an oracle), as well as that frame's role inventory, though it is not known which roles are instantiated in the given test instance. Role labeling is thus performed on an LU basis.

To evaluate frame labeling, we split the set of sentences by lemma and perform 5-fold cross-validation. Cross-validation splits for role labeling are done according to LU.

Unseen Data. To evaluate the performance of our system on unseen data,we simulate a situation in which individual LUs are unseen; specifically, we assume that the frame of a given LU has been seen before but not with the target lemma.[2] We also allow the case that a target lemma has been seen with a different frame. Note that while having seen the target frame before will help the model to select the correct frame, having seen the target lemma is not necessarily helpful, as it might lead the system to predict the incorrect seen frame rather than the correct but previously unseen frame.

To simulate the unseen condition for a given LU, all annotated sentences for that LU are removed from the training set and put into the test set. To test our hypothesis that the performance of correctly predicting a frame (and by extension also the roles) for an unseen LU depends on the frequency of the target frame after removing the LU, we computed the *inverse frequency* of each LU, i.e., the frame frequency summed over all other LUs with the same target frame, and sorted the set of LUs into three frequency bands based on their inverse frequency. Each band contains approximately the same number of LUs, subject to the constraint that LUs with the same inverse frequency are grouped together. A test set was then created by randomly selecting 10% of the LUs from each band, making sure that the test set contains each frame only once; the training set consists of all remaining LUs. Because this configuration does not allow proper cross-validation, instead five random training-test splits were created and tested.

5 Experiments

Automatic semantic analysis under the FrameNet approach is generally modeled as a two-part process: frame identification (Section 5.1) and role assignment (Section 5.2). Having a frame label for an instance's target lemma is a prerequisite to role assignment, as there is a distinct inventory of possible role labels for the semantic arguments of any given frame.[3] We evaluate our model independently on the two component tasks and then perform a preliminary evaluation on the complete semantic analysis task, taking a pipeline approach.

Features. The model uses nearly the same feature set for both prediction tasks, with a few exceptions. Table 1 shows which features are used for each task and whether the feature values are gold-standard or predicted.

	FramePred	RolePred	Pipeline
target lemma	G	G	G
target pos	G	G	G
# args	A	G	A
arg head	A	A*	A
arg head POS	A	A*	A
syn pattern	A	G	A
WordNet props	A	A	A
frame label	-	G	M
role types	-	M/G	M

Table 1: Features used for each task. **G**: gold-standard feature values; **A**: automatically-obtained feature values; **A***: automatically-obtained feature values based on gold-standard input; **M**: feature values predicted by our model

Values for automatically-obtained argument-related features are extracted from a metafeature representation produced by the frame assignment component of the Shalmaneser SRL system (Erk and Padó, 2006). The automatic syntactic patterns are then computed by aligning arguments with the text and replacing the arguments with their phrase-level syntactic categories.

WordNet features are extracted for each noun and verb in the lexicon. First, all hypernyms are extracted for the first sense of the word. In addition, one member from each hypernym synset is added to the list of properties for the lexical item.

Baselines and reporting. For each task we calculate an item baseline based on the number of possible outcomes. In the case of frame identification, the baseline reflects the number of frames a target predicate can participate in. If an LU exists in the frame dictionary, the number of possible outcomes is equal to the number of potential frame labels in the dictionary; if it does not, the denominator will be the total number of frame labels observed in the training data. For role labeling, the baseline reflects the number of roles available for labeling a given argument. Again, lemmas appearing in the frame-role dictionary have fewer possible labels. The baselines reported in Table 2 and Table 3 are the respective averages of all item baselines across different data sets.

Because our clustering algorithm is incremental and each training instance is processed only once, the model's performance in each task depends on the order of input items in the training set. In practice, though, no significant difference was observed across different cross-validation folds.

[2] This is in line with previous research on SRL for unseen data; creating or inducing entirely novel frames is beyond the reach of any current SRL system.

[3] Some role names appear in multiple frames, but they cannot necessarily be assumed to be semantically equivalent.

Frame Prediction		
Seen Data	Unseen Data	Baseline
88.32	88.76	87.09

Table 2: Accuracy of frame predictions for seen and unseen data, five-fold cross-validation.

Also, in the case of unseen data sets, no significant difference was observed across different frequency bands. Therefore, in the following sections, the reported results are averaged over all three frequency bands (as well as over all cross-validation folds).

5.1 Frame identification

For frame identification, we assume that the target lemma has been previously identified, and the model's predictions are constrained by a per-lemma frame dictionary built from FrameNet. This dictionary contains *all* LUs defined in FrameNet, so constraining the model with the frame dictionary is not equivalent to constraining the model to the LUs seen in training. This latter constraint is responsible for some of the coverage problems faced by other supervised models, so relaxing this constraint helps our model.

Results. The results for frame prediction on both seen and unseen data appear in Table 2. The high baseline figure reflects the fact that many lemmas in FrameNet appear with only a single frame. In combination with the frame dictionary, then, getting these right is a trivial matter. Nonetheless, our model improves on the baseline for both the seen and the unseen case. The latter is particularly positive as it means that we are able to infer the frame even for unseen LUs.

5.2 Role assignment

In a complete, end-to-end semantic role labeling system, role assignment involves both determining the span of the semantic arguments and assigning role labels to them. As our focus in this paper is the clustering model, we do not evaluate on the argument identification task, but rather assume gold-standard argument spans as input to role assignment. Having perfect argument spans greatly reduces the noisiness of both the argument head features and the syntactic pattern, at the same time improving the quality of the extracted WordNet features. Of course, assuming perfect input to role assignment is unrealistic for any real-world setting; thus we briefly report results on executing the entire SRL pipeline in Section 5.3.

Role Prediction		
	Seen Data	Unseen Data
no types	60.00	46.31
predicted types	67.00	46.29
gold types	74.84	73.65
Baseline	11.95	

Table 3: Accuracy of role assignment for seen and unseen data, five-fold cross-validation, with and without semantic types for roles.

The model's role label predictions are constrained using a frame-role dictionary extracted from FrameNet. For each individual instance, the set of available role labels is restricted to those defined for the frame assigned to the target lemma.

Predicted role types. As an additional feature for role assignment, we use semantic types on role fillers, as given in FrameNet. For example, for the frame COGITATION, the filler of the COGNIZER role must be a *Sentient* entity. Most types correspond to one or more WordNet synsets (Ruppenhofer et al., 2006). Unlike role names, these semantic types are not specific to frames, but rather shared across the lexicon.

In theory, these semantic types should be a powerful feature for assigning role labels. However, gold-standard semantic types are available for only a small part of the frame-specific roles defined in FrameNet. Though some previous work has used these semantic types to generalize over roles (Matsubayashi et al., 2009), no system so far has predicted role types to fill those gaps. To address this particular coverage problem, we first train a model on the available role types, predict values for all role types in the test data, and incorporate the predicted types as a novel feature for role assignment.

Results. Results for role assignment appear in Table 3. All results improve on the baseline. Unsurprisingly, gold standard role types lead to the largest performance gain. However, it can be seen that even when the role types are first predicted automatically, noticeable performance gains can be obtained compared to not using type information at all, at least for seen data. For unseen data automatically inferred type information does not help, possibly because the type prediction for LUs not seen in the training data is too noisy. Predictably, the results are lower for unseen data than for seen LUs, however, the model degenerates gracefully

Complete analysis		
	no types	predict types
Seen Data	41.80	45.76

Table 4: Performance on role prediction as a pipeline task, seen data only, five-fold cross-validation.

and is still able to correctly label almost every second argument for unseen LUs.

5.3 Complete semantic analysis

To evaluate our model's performance on complete semantic analysis, we use a pipeline approach: frame prediction, role type prediction, and role assignment. For all but the first task, predictions from the prior stage of analysis are fed into the model for the next. The only types of oracle information the model has access to are the target lemma and its part of speech tag, and the frame and role dictionaries described above.

Our results on seen data are in the same neighbourhood as the state-of-the-art. For example, the SEMAFOR system (Das et al., 2010) is reported to reach an F1 score of 46.00 for full parsing using oracle targets.

6 Conclusion

We present a Bayesian clustering and prediction model for FrameNet semantic role labeling. The proposed model is capable of generalizing its knowledge of similar frame instances to novel cases and is particularly competent in handling previously unseen data. Our results show that the model performs much better than chance in assigning semantic roles to arguments in an instance of a lexical unit which has not been seen in the training data. Also, the performance of the model for frame prediction on test sets of unseen data is as good as its performance on seen data.

We also propose a novel strategy which significantly improves the accuracy of SRL for seen data: we use all other features from an annotated instance to predict the most probable *role type*, and then use the predicted role type as an additional feature for predicting the semantic role.

Although we do not improve on state of the art results for frame prediction or role assignment, our model offers better coverage than existing models. In the future, we plan to improve the performance of our model by exploring the contribution of additional features (such as word classes and dependency relations between the arguments and the predicate), and to evaluate our model on additional data sets such as SemEval 2007.

Acknowledgments

This research has been funded by the German Research Foundation (DFG) under the MMCI Cluster of Excellence.

References

A. Alishahi, S. Stevenson. 2010. A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, 25(1):50–93.

U. Baldewein, K. Erk, S. Padó, D. Prescher. 2004. Semantic role labelling with similarity-based generalization using em-based clustering. In *Proc of Senseval-04*.

A. Burchardt, K. Erk, A. Frank. 2005. A WordNet detour to FrameNet. In B. Fisseni, H.-C. Schmitz, B. Schröder, P. Wagner, eds., *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, volume 8.

D. D. Cao, D. Croce, M. Pennacchiotti, R. Basili. 2008. Combining word sense and usage for modeling frame semantics. In *Proceedings of STEP-08*.

D. Das, N. Schneider, D. Chen, N. A. Smith. 2010. Probabilistic frame-semantic parsing. In *Proc of NAACL-HLT-10*.

K. Erk, S. Padó. 2006. Shalmaneser – a toolchain for shallow semantic parsing. In *Proc of LREC-06*.

H. Fürstenau, M. Lapata. 2009. Semi-supervised semantic role labeling. In *Proc of EACL-09*.

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

R. Johansson, P. Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proc of the Wkshp on Building Frame-semantic Resources for Scandinavian & Baltic Languages, NODALIDA*.

Y. Matsubayashi, N. Okazaki, J. Tsujii. 2009. A comparative study on generalization of semantic roles in framenet. In *Proc of ACL-09*.

S. Padó, M. Pennacchiotti, C. Sporleder. 2008. Semantic role assignment for event nominalisations by leveraging verbal data. In *Proc of Coling-08*.

A. Palmer, C. Sporleder. 2010. Evaluating FrameNet-style semantic parsing: The role of coverage gaps in FrameNet. In *Proc of COLING-10*.

M. Palmer, D. Gildea, P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–105.

M. Pennacchiotti, D. D. Cao, R. Basili, D. Croce, M. Roth. 2008. Automatic induction of FrameNet lexical units. In *Proc of EMNLP-08*.

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, J. Scheffczyk. 2006. FrameNet II: Extended Theory and Practice.