Leveraging VerbNet to build Corpus-Specific Verb Clusters

Daniel W Peterson and **Jordan Boyd-Graber** and **Martha Palmer**
University of Colorado
{daniel.w.peterson,jordan.boyd.graber,martha.palmer}@colorado.edu

Daisuke Kawhara
Kyoto University, JP
dk@i.kyoto-u.ac.jp

Abstract

In this paper, we aim to close the gap from extensive, human-built semantic resources and corpus-driven unsupervised models. The particular resource explored here is VerbNet, whose organizing principle is that semantics and syntax are linked. To capture patterns of usage that can augment knowledge resources like VerbNet, we expand a Dirichlet process mixture model to predict a VerbNet class for each sense of each verb, allowing us to incorporate annotated VerbNet data to guide the clustering process. The resulting clusters align more closely to hand-curated syntactic/semantic groupings than any previous models, and can be adapted to new domains since they require only corpus counts.

1 Introduction

In this paper, we aim to close the gap from extensive, human-built semantic resources and corpus-driven unsupervised models. The work done by linguists over years of effort has been validated by the scientific community, and promises real traction on the fuzzy problem of deriving meaning from words. However, lack of coverage and adaptability currently limit the usefulness of this work.

The particular resource explored here is VerbNet (Kipper-Schuler, 2005), a semantic resource built upon the foundation of verb classes by Levin (1993). Levin's verb classes are built on the hypothesis that syntax and semantics are fundamentally linked. The semantics of a verb affect the allowable syntactic constructions involving that verb, creating regularities in language to which speakers are extremely sensitive. It follows that grouping verbs by allowable syntactic realizations leads from syntax to meaningful semantic groupings. This seed grew into VerbNet, a process which involved dozens of linguists and a decade of work, making careful decisions about the allowable syntactic frames for various verb senses, informed by text examples.

VerbNet is useful for semantic role labeling and related tasks (Giuglea and Moschitti, 2006; Yi, 2007; Yi et al., 2007; Merlo and van der Plas, 2009; Kshirsagar et al., 2014), but its widespread use is limited by coverage. Not all verbs have a VerbNet class, and some polysemous verbs have important senses unaccounted for. In addition, VerbNet is not easily adaptable to domain-specific corpora, so these omissions may be more prominent outside of the general-purpose corpora and linguistic intuition used in its construction. Its great strength is also its downfall: adding new verbs, new senses, and new classes requires trained linguists - at least, to preserve the integrity of the resource.

According to Levin's hypothesis, knowing the set of allowable syntactic patterns for a verb sense is sufficient to make meaningful semantic classifications. Large-scale corpora provide an extremely comprehensive picture of the possible syntactic realizations for any particular verb. With enough data in the training set, even infrequent verbs have sufficient data to support learning. Kawahara et al. (2014) showed that, using a Dirichlet Process Mixture Model (DPMM), a VerbNet-like clustering of verb senses can be built from counts of syntactic features.

We develop a model to extend VerbNet, using a large corpus with machine-annotated dependencies. We build on prior work by adding partial supervision from VerbNet, treating VerbNet classes as additional latent variables. The resulting clusters are more similar to the evaluation set, and each cluster in the DPMM predicts its VerbNet class distribution naturally. Because the technique is data-driven, it is easily adaptable to domain-specific corpora.

*Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*, pages 102–107,
Berlin, Germany, August 11-12, 2016.

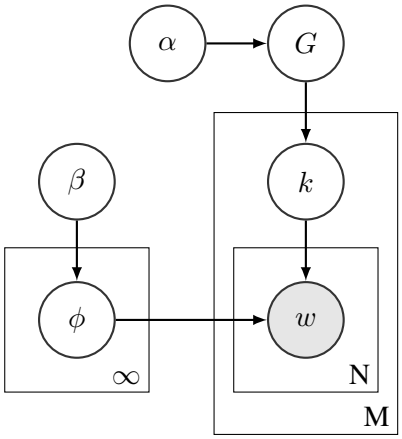

Figure 1: The DPMM used in Kawahara et al. (2014) for clustering verb senses. M is the number of verb senses, and N is the sum total of slot counts for that verb sense.

2 Prior Work

Parisien and Stevenson (2011) and Kawahara et al. (2014) showed distinct ways of applying the Hierarchical Dirichlet Process (Teh et al., 2006) to uncover the latent clusters from cluster examples. The latter used significantly larger corpora, and explicitly separated verb sense induction from the syntactic/semantic clustering, which allowed more fine-grained control of each step.

In Kawahara et al. (2014), two identical DPMM's were used. The first clustered verb instances into senses, and one such model was trained for each verb. These verb-sense clusters are available publicly, and are used unmodified in this paper. The second DPMM clusters verb senses into VerbNet-like clusters of verbs. The result is a resource that, like Verbnet, inherently captures the inherent polysemy of verbs. We focus our improvements on this second step, and try to derive verb clusters that more closely align to VerbNet.

2.1 Dirichlet Process Mixture Models

The DPMM used in Kawahara et al. (2014) is shown in Figure 1. The clusters are drawn from a Dirichlet Process with hyperparameter α and base distribution G. The Dirichlet process prior creates a clustering effect described by the Chinese Restaurant Process. Each cluster is chosen proportionally to the number of elements it already contains, i.e.

$$P(k|\alpha, C_k(*)) \propto \begin{cases} C_k(*), & \text{if } C_k(*) > 0 \\ \alpha, & \text{if } k = k_{new}, \end{cases} \quad (1)$$

where $C_k(*)$ is the count of clustered items already in cluster k.

Each cluster k has an associated multinomial distribution over vocabulary items (e.g. slot:token pairs), ϕ_k, which is drawn from G, a Dirichlet distribution of the same size as the vocabulary, parameterized by a constant β. Because the Dirichlet is the multinomial's conjugate prior, we can actually integrate out ϕ_k analytically, given counts of vocabulary items drawn from ϕ_k. For a particular vocabulary item w, we compute

$$P(w|\phi_k, \beta) = \frac{C_k(w) + \beta}{C_k(*) + |V|\beta}, \quad (2)$$

where $C_k(w)$ is the number of times w has been drawn from ϕ_k, $C_k(*) = \sum_i C_k(i)$, and $|V|$ is the size of the vocabulary.

When assigning a verb instance to a sense, a single instance may have multiple syntactic arguments w. Using Bayes's law, we update each assignment iteratively using Gibbs sampling, using equations (1) and (2), according to

$$P(k|\alpha, C_k(*), \phi_k, \beta) \propto P(k|\alpha, C_k(*)) \prod_w P(w|\phi_k, \beta). \quad (3)$$

$\beta < 1$ encourages the clusters to have a sparse representation in the vocabulary space. $\alpha = 1$ is a typical choice, and encourages a small number of clusters to be used.

2.2 Step-wise Verb Cluster Creation

By separating the verb sense induction and the clustering of verb senses, the features can be optimized for the distinct tasks. According to (Kawahara et al., 2014), the best features for inducing verb classes are joint slot:token pairs. For the verb clustering task, slot features which ignore the lexical items were the most effective. This aligns with Levin's hypothesis of diathesis alternations - the syntactic contexts are sufficient for the clustering.

In this paper, we re-create the second stage clustering with the same features, but add supervision. Supervised Topic Modeling (Mimno and McCallum, 2008; Ramage et al., 2009) builds on the Bayesian framework by adding, for each item, a

prediction about a variable of interest, which is observed at least some of the time. This encourages the topics to be useful at predicting a supervised signal, as well as coherent as topics. We do not have explicit knowledge of VerbNet class for any of the first-level DPMM's verb senses, so our supervision is informed only at the level of the verb.

3 Supervised DPMM

Adding supervision to the DPMM is fairly straightforward: at each step, we sample both a mixture component k and a VerbNet class y. For this, we assign each cluster (mixture component) a unique distribution ρ over VerbNet classes, drawn from a fixed-size Dirichlet prior with parameter γ. As before, this allows us to estimate the likelihood of a VerbNet class y knowing only the counts of assigned senses, $C_k(y)$, for each y, as

$$P(y|\rho_k, \gamma) = \frac{C_k(y) + \gamma}{C_k(*) + |S|\gamma}, \tag{4}$$

where $|S|$ is the number of classes in the supervision.

The likelihood of choosing a class for a particular verb requires us to form an estimate of that verb's probability of joining a particular VerbNet class. We initialize η from SemLink, as $\eta(y) = \omega * C_v^{SL}(y) + \delta$, for fixed constants ω and δ, and with $C_v^{SL}(y)$ as the count, in SemLink, of times verb v was assigned to VerbNet class y. We then draw a verb-specific distribution θ over VerbNet classes, from a Dirichlet with parameters η, so that η acts as pseudo-counts, steering θ to give high weight to VerbNet classes aligned with SemLink for each verb. We compute

$$P(y|\theta, \eta) = \frac{C_v(y) + \eta(y)}{C_v(*) + \sum \eta}, \tag{5}$$

where $C_v(y)$ is the number of times verb v is assigned to VerbNet class y by our model.

We sample the VerbNet class for a verb sense as a product of experts (Hinton, 2002), the θ_v for the verb v, and ρ_k for the assigned cluster k. This encourages alignment between the VerbNet classes observed in SemLink and the VerbNet classes predicted by the clusters, and is computationally straightforward. We simply compute

$$P(y|\rho_k, \gamma, \theta_v, \eta) \propto P(y|\rho_k, \gamma)P(y|\theta_v, \eta). \tag{6}$$

Sampling a cluster for a verb sense now depends on the VerbNet class y,

$$\begin{aligned} P(k|y, \alpha, \phi_k, \beta, \rho_k, \gamma, \theta_v, \eta) \propto \\ \Big(P(k|\alpha, C_k(*)) \times \\ P(y|\rho_k, \gamma, \theta_v, \eta) \times \\ \prod_w P(w|\phi_k, \beta)\Big). \end{aligned} \tag{7}$$

We then update y based on Equation 6, and then resample for the next batch.

The supervised process is depicted in Figure 2. In brief, we know for each verb an η, a given by counts from SemLink, which we use as a prior for θ. We sample, in addition to the cluster label k, a VerbNet class y, which depends on θ and ρ, where ρ is the distribution over VerbNet classes in cluster k. ρ is drawn from a Dirichlet distribution paramaterized by $\gamma < 1$, encouraging each cluster to have a sparse distribution over VerbNet classes. Because y depends on both θ and ρ, the clusters are encouraged to align with VerbNet classes.

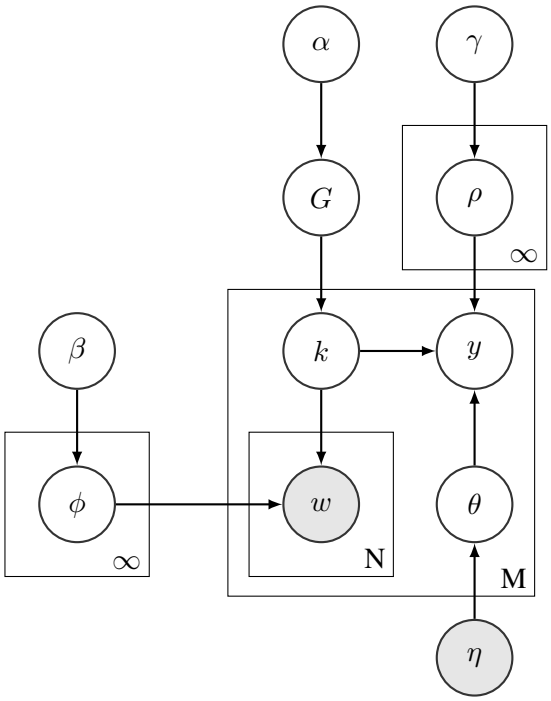

Figure 2: The Supervised DPMM used in this work for clustering verb senses. M is the number of verb senses, and N is the sum total of slot counts for that verb sense. θ is initialized to reflect the VerbNet class preferences for each verb, when they are known.

3.1 Modeling Choices

When incorporating supervision, the more direct method of downstream sampling of the VerbNet class may be preferred to using a prior. However, the verb senses are generated through a DPMM, and we do not have a gold-label assignment of VerbNet classes to each sense. Instead, we estimate, for each verb in VerbNet, a distribution θ describing the likelihood a verb will participate in a particular class, using counts from SemLink.

When sampling a cluster for a verb sense with a verb in VerbNet, we sample y from a product of experts. We cannot incorporate θ as a prior when sampling y, because we have multiple verbs, with distinct distributions $\theta_{v_1}, \theta_{v_2}, \ldots$.

Because the product-of-experts is a discrete probability distribution, it is easy to marginalize out this variable when sampling k, using

$$P(k|\alpha, \phi_k, \beta, \rho_k, \gamma, \theta) \propto \sum_y P(k|y, \alpha, \phi_k, \beta, \rho_k, \gamma, \theta_v, \eta). \quad (8)$$

Either way, once a cluster is selected, we should update the ρ and θ. So, once a cluster is selected, we still sample a discrete y. We compare performance for sampling k with assigned y and with marginalized y.

When incorporating supervision, we flatten VerbNet, using only the top-level categories, simplifying the selection process for y. In Kawahara et al. (2014), slot features were most effective features at producing a VerbNet-like structure; we follow suit.

4 Results

For evaluation, we compare using the same dataset and metrics as Kawahara et al. (2014). There, the authors use the polysemous verb classes of Korhonen et al. (2003), a subset of frequent polysemous verbs. This makes the test set a sort of mini-VerbNet, suitable for evaluation. They also define a normalized modified purity and normalized inverse purity for evaluation, explained below.

The standard purity of a hard clustering averages, for each cluster's majority gold standard class, the percentage of clustered items of that class. Because the clustering is polysemous, a typical automatically-induced cluster K will contain only some senses of the verbs. We take this partial membership into account when deciding the cluster's majority class. We define $c_{iv} \in [0, 1]$ as the proportion of instances of verb v grouped into cluster K_i. We also treat induced clusters containing only one verb sense as errors, rather than treating them as clusters of perfect purity. Therefore, the normalized modified purity (nmPU), with respect to the gold standard clusters G, is,

$$\text{nmPU} = \frac{1}{N} \sum_{i \text{ s.t. } |K_i|>1} \max_j \delta_{K_i}(K_i \cap G_j), \quad (9)$$

where

$$\delta_{K_i}(K_i \cap G_j) = \sum_{v \in K_i \cap G_j} c_{iv}. \quad (10)$$

This nmPU is analagous to clustering precision: it measures, on average, how well the clustering avoids matching items that should not be clustered. We also define a recall analogue, the normalized inverse purity (niPU), as,

$$\text{niPU} = \frac{1}{N} \sum_j \max_i \delta_{G_j}(K_i \cap G_j). \quad (11)$$

This measures how well each gold standard cluster is recovered. We report each metric, and the F1 score combining them, to compare the clustering accuracy with respect to the gold standard G.

We use the clustering from Kawahara et al. (2014) as a baseline for comparison. However, for evaluation, the authors only clustered senses of verbs in the evaluation set. Since we would like to test the effectiveness of adding supervision, we treat all verbs in the evaluation set as unsupervised, with no initialization of θ. Therefore, to compare apples-to-apples, we calculate the nPU, niPU, and F1 of the Kawahara et al. (2014) full clustering against the evaluation set. Our model also computes the full clustering, but with supervision for known verbs (other than the evaluation set).

Parameters were selected using a grid search, and cross-validation. The results are summarized in Table 1, comparing the unsupervised DPMM baseline (**DPMM**) to the supervised DPMM (**SDPMM**), and the supervised DPMM sampling k with y marginalized out (**mSDPMM**).

5 Comparison of Produced Clusters

The supervised sampling scheme produces fewer clusters than the unsupervised baseline. This is in

Model	Example Clusters	
Gold	**push** (0.20), **pull** (0.17)	**give** (1.0), **lend** (1.0), **generate** (0.33), **allow** (0.25), **pull** (0.17), **pour** (0.17)
DPMM	**push** (0.40), **drag** (0.27), **pull** (0.08)	**lend** (0.30), **give** (0.13),
SDPMM	**drag** (0.87), **push** (0.43), **pull** (0.42), **pour** (0.39), **drop** (0.31), **force** (0.09)	**give** (0.82), **pour** (0.02), **ship** (0.002)

Table 2: Example clusters from the evaluation dataset (**Gold**), and along with the most-aligned clusters from the unsupervised baseline (**DPMM**) and our semi-supervised clustering scheme (**SDPMM**). Weights given in parentheses describe the total proportion of verb instances assigned to each cluster.

Model	nmPU	niPU	F1	N
DPMM	**55.72**	60.33	57.93	522
SDPMM	51.00	**75.71**	**60.95**	122
mSDPMM	51.04	75.00	60.74	129

Table 1: Clustering accuracy on verbs in the Korhonen et al. (2003) dataset. N is the number of clusters spanned by the evaluation set.

part because it produces fewer "singleton" clusters, containing only one verb sense from the evaluation set. The SDPMM produces only 16% singleton clusters, compared with 34% of singleton clusters from the unsupervised DPMM.

The supervised clusters also tend to cluster more of the senses of each verb into the same cluster. The predominant SDPMM cluster for a verb, which has the highest percentage of a verb's total instances, tends to have 224% the number of instances as the predominant unsupervised DPMM cluster. This tendency does not prevent verbs being assigned multiple clusters, however. On average, the supervised clustering uses 30% fewer clusters for each verb, a smaller reduction than the 70% overall drop in the number of clusters.

A few example clusters are presented in Table 2.

6 Conclusions and Future Directions

The supervision tends to encourage a smaller number of clusters, so the precision-like metric, nmPU, is lower, but the recall-like metric, niPU, is much higher. Marginalizing out the variable y when sampling k does not make an appreciable difference to the F1 score. Swapping out the Dirichlet process for a Pitman-Yor process may bring finer control over the number of clusters.

We have expanded the work in Kawahara et al. (2014) by explicitly modeling a VerbNet class for each verb sense, drawn from a product of experts based on the cluster and verb. This allowed us to leverage data from SemLink with VerbNet annotation, to produce a higher-quality clustering. It also allows us to describe each cluster in terms of alignment to VerbNet classes. Both of these improvements bring us closer to extending VerbNet's usefulness, using only automated dependency parses of corpora. We may speculate, and should test, whether the improved verb clusters will prove useful in end-to-end semantic tasks.

References

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via framenet, verbnet and propbank. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*.

Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Daisuke Kawahara, Daniel W. Peterson, and Martha Palmer. 2014. A step-wise usage-based method for inducing polysemy-aware verb classes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014)*.

Karin Kipper-Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Anna Korhonen, Yuval Krymolowski, and Zvika Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*, pages 64–71.

Meghana Kshirsagar, Nathan Schneider, and Chris Dyer. 2014. Leveraging heterogeneous data sources for relational semantic parsing. In *Proceedings of ACL 2014 Workshop on Semantic Parsing*.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. The University of Chicago Press.

Paola Merlo and Lonneke van der Plas. 2009. Abstraction and generalisation in semantic role labels: Propbank, verbnet or both? In *Proceedings of IJCNLP/ACL 2009*.

David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*.

Christopher Parisien and Suzanne Stevenson. 2011. Generalizing between form and meaning using learned verb classes. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci2011)*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).

Szu-ting Yi, Edward Loper, and Martha Palmer. 2007. Can semantic roles generalize across genres? In *Proceedings of NAACL HLT 2007*.

Szu-ting Yi. 2007. *Automatic Semantic Role Labeling*. Ph.D. thesis, University of Pennsylvania.