Verb Class Induction with Partial Supervision

Daniel Peterson
Oracle Labs
Burlington, MA 01803
daniel.peterson@oracle.com

Susan W. Brown, Martha Palmer
University of Colorado
Boulder, CO 80309
{susan.brown, martha.palmer}@colorado.edu

Abstract

Dirichlet-multinomial (D-M) mixtures like latent Dirichlet allocation (LDA) are widely used for both topic modeling and clustering. Prior work on constructing Levin-style semantic verb clusters achieves state-of-the-art results using D-M mixtures for verb sense induction and clustering. We add a bias toward known clusters by explicitly labeling a small number of observations with their correct VerbNet class. We demonstrate that this partial supervision guides the resulting clusters effectively, improving the recovery of both labeled and unlabeled classes by 16%, for a joint 12% absolute improvement in F1 score compared to clustering without supervision. The resulting clusters are also more semantically coherent. Although the technical change is minor, it produces a large effect, with important practical consequences for supervised topic modeling in general.

Introduction

VerbNet (Kipper-Schuler 2005; Kipper et al. 2006) is a useful semantic resource but faces coverage issues. It is organized on principles outlined by (Levin 1993), who demonstrated a significant link between verb semantics and the allowed syntactic structures for those verbs. VerbNet has been greatly expanded by linguists over the years (Bonial, Stowe, and Palmer 2013a), adding coverage while adhering to its organizing principles. VerbNet now contains 270 classes and 5300 entries, and its granularity is relatively coarse while providing clear semantic groupings. VerbNet class annotations have proven useful for tasks like semantic role labeling (Giuglea and Moschitti 2006; Hartmann, Eckle-Kohler, and Gurevych 2016), which is an intermediate task used for systems like information extraction and question answering (Shen and Lapata 2007; Christensen et al. 2010; Moreda et al. 2011). Its success in supporting NLP tasks has led to the creation of similar resources in other languages, such as Urdu (Hautli-Janisz, King, and Ramchand 2015), French (Pradet, Danlos, and De Chalendar 2014), Basque (Aldezabal et al. 2010) and Arabic (Mousser 2010).

VerbNet benefits from linguistic theory and careful curation, but the time and effort required to create such a resource manually is substantial. A means of automatically inducing verb sense clusters could aid in the creation of new VerbNets for low-resource languages. In addition, automatic clustering could improve the coverage of the English VerbNet and similar resources in other languages. The English VerbNet is still missing many verbs and some important senses of common verbs. As the resource grows, the task of adding new elements becomes harder, and the long-tail nature of language use means there are many thousands of infrequent verbs, verb senses, or verb-particle constructions that may require special handling. Interest in expanding VerbNet algorithmically and generating VerbNets for other languages has led to a rich history of algorithms for generating Levin-style verb clusters (Kipper et al. 2000; Im Walde 2000; Brew and Schulte im Walde 2002; Korhonen, Krymolowski, and Marx 2003; Lapata and Brew 2004; Li and Brew 2008; Sun and Korhonen 2011; Majewska et al. 2018).

Probabilistic graphical models have shown the most promise creating VerbNet-like clusterings from minimally-processed corpus data (Kawahara, Peterson, and Palmer 2014; Peterson and Palmer 2018). In these approaches, each sentence in the corpus is used as context to assign a cluster label to the root verb of the sentence. Clustering approaches have the advantage that they can discover new verb senses to add to existing VerbNet classes, but can also propose new VerbNet classes. However, these automatically-induced clusters have been noisy, and have not yet been used by VerbNet annotators.

In this paper, we develop a simple, computationally efficient technique to bias the sampling procedure toward discovering a known subset of VerbNet classes, on a limited set of annotations. This builds on state-of-the-art probabilistic clustering approaches for creating VerbNet clusters, which use Dirichlet-multinomial (D-M) mixtures for both sense induction and clustering (Kawahara, Peterson, and Palmer 2014; Peterson and Palmer 2018). We have a small number of sentences labeled with VerbNet senses, and we bias our sampling by observing these labeled sentences in their correct clusters at initialization, and not re-sampling these initial assignments. The sampling procedure for all other tokens in the corpus proceeds as normal, so this initialization causes negligible impact on the running time of the algorithm, and allows the sense induction and clustering for all other verbs to proceed unchanged.

This method uses a small amount of labeled data, in com-

bination with state-of-the-art techniques for creating Levin-style clusters, and is extremely practical. Our aim is to multiply the impact of annotation, and while our experiments focus on English, we believe this work can assist the creators of VerbNet-like resources in multiple languages (Mousser 2010; Danlos, Nakamura, and Pradet 2014; Pradet, Danlos, and De Chalendar 2014; Scarton, Duran, and Aluísio 2014; Estarrona et al. 2015; Hautli-Janisz, King, and Ramchand 2015; Vulić, Mrkšić, and Korhonen 2017).

Our main contribution is the demonstration that this simple supervision technique aids in the recovery of VerbNet classes. We split the labeled instances into training and test portions, and show that including partial supervision improves the recovery of VerbNet classes by 10% F1 on the test set, and qualitatively improves the semantic coherence of the resulting clusters. We also demonstrate that augmenting the model's vocabulary improves VerbNet clustering, yielding a 5% increase in F1 score over the prior state-of-the art, on top of the supervision.

Background

The probabilistic models underlying this work are most similar to latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003), which is widely used for generating semantically-coherent topics out of a large corpus. Each topic is a unigram language model (i.e., a multinomial probability distribution over the vocabulary) that describes a set of frequent co-occurring terms in the corpus. These multinomials are drawn from a Dirichlet distribution, which encourages the weight to be concentrated on a small number of words. In the generative model, each document is represented by a separate multinomial distribution over topics, also drawn from a Dirichlet. Each word is generated by drawing a latent topic assignment from the document's distribution, and then a word from the corresponding topic distribution. The Dirichlet priors provide probabilistic constraints on the topic assignments, encouraging "not too many" topics per document and "not too many" words per topic, but without requiring hard limits for any particular document or topic. Fitting this model to the data, then, is the act of learning the latent topic assignments for each word in the corpus, in a way that balances these priors against the observations.

There is a critical link between topic modeling and clustering that we exploit in this work. Since each word in the corpus is explicitly assigned a single topic, we can treat the topic assignments as a clustering of the corpus, at the word level. Topics from LDA are widely used because they are semantically coherent, and our aim is to produce semantic verb clusters. We use D-M mixtures or closely-related Dirichlet Process mixtures to capture semantically coherent clusters from our data.

Our work builds on a successful probabilistic framework for generating VerbNet-like clusters from a dependency-parsed corpus. The step-wise method (Kawahara, Peterson, and Palmer 2014) first captures verb polysemy by clustering sentences with different meanings into different senses (e.g., "John entered the room" is semantically different from "Jane entered the military"). Each verb has its sentences clustered into senses independently, learning topic distributions over a vocabulary of `slot:token` pairs (e.g., "John entered the room" has features `subject:John` and `direct_object:room`). A second mixture model clusters learned senses across verbs (e.g., linking the "John entered a room" sense to "approach" and the "Jane entered the military" sense to "join"). The clustering step uses topics over `slot` features (e.g., both example sentences have the features `subject` and `direct_object`), because abstracting away from the tokens helps recover VerbNet-like classes. D-M mixtures work for both steps of this process and produced state-of-the-art results[1]. We adopt these same sets of features to represent sentences throughout this work, but the model we employ is computationally simpler.

A more recent paper proposed a single-step, joint sense induction and clustering framework that is nearly identical to LDA, and achieved higher clustering accuracy than the step-wise process (Peterson and Palmer 2018). At its most basic, the single-step process is running LDA on a corpus in which each document is the collection of sentences with the same verb across the corpus, and with the constraint that each sentence must be assigned to a single topic. Sharing topics across verbs, rather than developing topics in isolation, produces higher-quality verb senses, but also produces a semantic clustering of verbs. Each topic is a verb cluster, and each topic used by a given verb corresponds to a distinct sense of that verb. Because the single-step framework doesn't require sampling the corpus multiple times, it has a lower computational burden, and because it requires only minimal modification to the LDA algorithm, it allows us to exploit an existing, fast, and distributed implementation of LDA for our research. This single-step model is the base model for our work. Our key contribution is efficiently adding partial supervision to this model, allowing some additional control over the output clustering and improving accuracy.

Adding partial supervision to probabilistic clustering techniques can help recover the desired clusters. (Peterson et al. 2016) added VerbNet class preferences to the clustering step of the step-wise process, but did not directly use labeled sentences to guide the clustering. An automatically-acquired sense may contain sentences from multiple VerbNet classes, so there is no straightforward way to extrapolate from sentence labels to labels for the senses. The step-wise construction allows senses to be learned at a different granularity of features, but also places a barrier between the labeled data and the clustering objective. Instead, during sampling, each sense was assigned both a cluster and a VerbNet class. Data about the VerbNet class preferences for a verb in the supervised set were included as biases in the sampling of a class for the sense. This added both an additional factor to the sampling of clusters and an additional variable to sample, resulting in a computationally intensive process. Although this prior work demonstrated the potential of partial supervision, it was based on the step-wise verb clustering process (Kawahara, Peterson, and Palmer 2014).

Our approach to adding supervision requires less compu-

[1] At the time of publication.

tational overhead, produces more consistent output clusters, and is more efficient, because it is based on the single-step clustering process (Peterson and Palmer 2018). The key to the efficiency is that, in the single-step process, the clusters are generated directly from the sentences, with no intermediate steps to obfuscate the labeling. In the step-wise framework for clustering, adding supervision required an additional sampled VerbNet class variable to the inner loop of the MCMC algorithm, because the sentences were abstracted from the senses. In the joint framework, the cluster assignments can be treated directly as VerbNet class assignments of the sentences, so we know the correct cluster assignment for any labeled class. This permits us to develop a scheme for semi-supervised clustering that builds from labeled sentences to higher-quality clusters.

Semi-Supervised Clustering with Direct Observations

To guide our topic model so that its learned topics closely match VerbNet, we explicitly observe some sentences that have a labeled VerbNet class. If our labels span C classes, we use a minimum of C topics, and assign each VerbNet class c_i to a topic k_i. When initializing our topic model, we normally assign each sentence to a random topic, and update the statistics for the Gibbs sampler, which repeatedly updates these topic assignments until convergence. Now, when we initialize our topic model, we also explicitly observe some labeled sentences and assign each sentence with VerbNet class c_i to topic k_i, and we leave this assignment fixed throughout sampling. All unlabeled data is treated normally, at initialization and during sampling.

This is a simple and straightforward means of guiding the clusters, but differs from prior work. Verbs with labeled sentences are biased to participate in the correct VerbNet classes, and the topics are biased to contain the vocabulary items corresponding to those same classes. We tune the weight of that bias by observing each labeled sentence w times, because there are orders of magnitude more unlabeled data than labeled. Once sufficient statistics are initialized, however, there are no further changes to the sampling algorithm.

Supervised LDA (Mcauliffe and Blei 2008) and DiscLDA (Lacoste-Julien, Sha, and Jordan 2009) both add a secondary classification task to the training objective, so that the topics assigned to each topic are effective features to classify the document according to a fixed label. Labeled LDA (Ramage et al. 2009) designates a single, known label to each topic, and allows documents to have multiple labels. The labels on the documents provide hard constraints on the available topics, so the learned topics conform to the label set, regardless of semantic interpretability. These techniques require accurate and complete document labels in order to be effective, and they limit the applicability to semi-supervised domains. They aren't really suitable for our task, because we know some verbs have senses not currently catalogued in VerbNet and many verbs are missing entirely. These gaps mean we do not have the required exhaustive labeling of our documents for supervised LDA.

There are several methods of including word cooccurrence knowledge or constraints to help ensure topics conform to user-specified constraints (Xie, Yang, and Xing 2015; Yang, Downey, and Boyd-Graber 2015; Hu et al. 2014; Andrzejewski, Zhu, and Craven 2009; Jagarlamudi, Daumé III, and Udupa 2012), that allow users to specify words that must or must not belong together, and in doing so guide the output of the model without exhaustive labeling of the documents. However, they require structural changes to the model that increase the computational burden during sampling. They are also unsuitable to our desired task, because our knowledge is given by labels on specific sentences, not pairwise vocabulary interactions.

Our implementation allows the user to specify partial information about VerbNet classes, to help the model conform to this prior knowledge without requiring a complete specification. Because it does not require any change to the training objective, it creates negligible computational burden. It uses the labeled examples we have, but allows the model room to discover novel classes and novel verb senses, as required to fit the unlabeled data. These are strong advantages over existing work, and our experiments demonstrate it is surprisingly effective.

Evaluation

Semlink (Bonial, Stowe, and Palmer 2013b) provides labels of VerbNet class for each sentence in the Penn Treebank's Wall Street Journal corpus (Marcus et al. 1994). These labels were used to evaluate the quality of sense induction and clustering in prior work, but they are also a potentially valuable resource to guide sense induction.

To test whether a small number of labels can improve the senses learned from LDA, we split this annotation into a training portion and a test portion. The split was designed to address two separate concerns. First, can partial supervision of a VerbNet class improve the recovery of that class from the topic model? Second, can supervision of some known classes aid the recovery of other classes? To address both these concerns, we first split the data by VerbNet class, using 2/3 of the classes as training (hereafter, C_1 denotes the set of classes in the training portion of the split) and 1/3 for testing (C_2). We then split by verb, keeping 2/3 for training and 1/3 for testing. We only use examples from the 141 most frequent verbs in Semlink, whose labeled sentences span 148 VerbNet classes. This training/test split produced 6400 sentences with known labels for training and 6500 for testing.

Our primary sources of data are Gigaword (Parker et al. 2011) and the Wall Street Journal sections of the Penn Treebank (Marcus et al. 1994), both licensed through the Linguistic Data Consortium. Gigaword is tokenized and dependency parsed automatically as a preprocessing step. Each "document" in LDA is the set of syntactic dependencies observed for a particular verb. The "words" in the document are either `slot` or `slot:token` observations. According to (Kawahara, Peterson, and Palmer 2014), the best features for inducing verb senses are joint `slot:token` pairs. For the verb clustering task, `slot` features that ignore the lexical items were the most effective. The best single-step model, empirically, uses both sets of vocabulary together,

effectively counting each token twice (once with and once without the corresponding lexical item). We only consider direct dependencies of the verbs and prepositional objects labeled with the observed preposition.

When training in the supervised setting, we include the 6400 sentences with known labels and assign each label to a particular topic. These assignments are never resampled, so throughout sampling, the supervised verb has some higher-than-random probability mass assigned to the designated topics, and the topics always have some higher-than-random probability mass assigned to the associated vocabulary items. Because Gigaword is much larger than our supervision set, we include a hyperparameter to increase the weight of these labeled instances. Effectively, we label the known sentences as though we'd seen them all many times.

Quantitative Evaluation Protocol

Once a model is trained, we assign each test sentence to its maximum a posteriori topic, and treat all sentences assigned the same topic as belonging to the same cluster. Each test sentence has a correct label, so we have a ground-truth clustering from this annotation. Following the conventions in the literature, we report standard clustering metrics.

The modified purity (*mPU*) is analogous to precision and measures how well the model distinguishes the different classes in the evaluation set, G. Each cluster K from the model is labeled with its majority class from G, and the purity of that cluster is the ratio of this majority class to its size. If a clustering K assigns singleton clusters, they are guaranteed to have perfect purity, so the modified purity removes these. In particular, the modified purity is the micro-average of these cluster purities, dividing the number of correctly-grouped sentences by the size of the evaluation set. If there are N sentences in the data set, and $C(c_1, c_2)$ counts the number of elements in a cluster c_1 that are also elements of c_2,

$$\text{mPU}(K, G) = \frac{1}{N} \sum_{c_1 \in K,\ |c_1|>1} \max_{c_2 \in G} C(c_1, c_2). \quad (1)$$

The inverse purity (*iPU*) is analogous to recall and measures how completely the clusters in the evaluation set are recovered.

$$\text{iPU}(K, G) = \frac{1}{N} \sum_{c_2 \in G} \max_{c_1 \in K} C(c_1, c_2). \quad (2)$$

There's no need to modify the inverse purity, because singleton clusters in the gold standard are correct, even though they will certainly be recovered fully. The harmonic mean of these two measures (*F1*) is a good measure of how closely the two clusterings align. All three scores are between 0 and 1, with 1 being a perfect recovery. We report them as percentages between 0 and 100.

Quantitative Evaluation Results

The Step-wise model splits sense induction and clustering into independent steps (Kawahara, Peterson, and Palmer 2014), performs on-par with the Joint model which learns senses and clusters simultaneously (Peterson and Palmer 2018). The Step-wise model uses both `slot:token` pairs and `slot` features as vocabulary, and using both sets of features on the Joint model (Joint + `slot`) significantly improves the Joint model results.

Adding partial supervision to these models significantly improves the clustering quality. Supervision in the Step-wise model (Peterson et al. 2016) dramatically boosts the mPU score of the clusters, improving absolute F1 by nearly 10%, and requires a significant increase in computational complexity. Adding supervision to the Joint model using our method significantly improves both mPU and iPU of the clusters, producing a nearly 12% absolute F1 score improvement without increasing computational complexity.

The Joint model with partial supervision, and using both `slot:token` and `slot` features, significantly outperforms all other models at recovering the clustering in our test set.

Adding supervision by biasing particular topics dramatically increases the consistency of the topics learned. In Tables 1 and 2, we report the mean and standard deviations of the scores across ten runs of each model. Joint + SS models have lower standard deviations, but they are also extremely consistent at recovering nearly the same clusters, run after run, for the seeded topics. Typically in topic modeling applications, different starting conditions produce different clusters, highlighting and obfuscating different themes. However, each seeded topic consistently produced the same clusters. This enhances the practicality of this technique for building VerbNet-style clusters, because adding supervision of new classes should have predictable effects despite the randomized nature of MCMC.

Qualitative Evaluation

VerbNet classes have been manually constructed based on the theory that syntactic patterns reflect semantic similarity. Each class comprises anywhere from 2 to over 100 related verb senses, although most classes have 10 to 30 verb senses. Distinctions between VerbNet classes can be subtle, occasionally even depending on a single syntactic alternation, such as whether the verbs appear in the ditransitive construction (e.g., John gave Mary a book vs. *John obtained Mary a book).

A qualitative analysis of our semi-supervised clusters shows that semantic similarities are being captured, but not yet at the fine distinctions of VerbNet classes. One of the topics clustered instances of *take, grab* and *seize,*. The topic was seeded with *take* sentences labeled with the Steal-10.5 class. As hypothesized, test sentences from the Steal class for *grab* and *seize* were correctly clustered here. However, all these verbs are polysemous, and sentences labeled with the class Obtain-13.5.2 for *grab* and *seize* were clustered here as well. These VerbNet classes are closely related, with some member verbs cross-listed between them. The main difference is the emphasis on the entity losing possession in Steal-10.5 but on the acquisition of something in Obtain-13.5.2, a subtlety the algorithm cannot yet make.

In addition, this cluster illustrates the need for careful balancing of seed sentences. The most numerous sentences gathered in this topic were *take* sentences labeled with

Model	mPU	iPU	F1
Step-wise	48.06 ± 5.00	54.00 ± 1.31	50.69 ± 2.60
Step-wise + SS	$\mathbf{67.12} \pm 1.88$	54.70 ± 1.17	60.26 ± 1.02
Joint	53.42 ± 0.91	43.83 ± 0.31	48.13 ± 0.75
Joint + SS	64.04 ± 0.38	54.03 ± 0.58	58.61 ± 0.45
Joint + `slot`	59.08 ± 0.39	46.49 ± 1.64	52.02 ± 1.09
Joint + `slot` + SS	65.73 ± 0.49	$\mathbf{62.29} \pm 0.74$	$\mathbf{63.96} \pm 0.58$

Table 1: Clustering accuracy on the complete test set, for various models. The Step-wise model with partial supervision (+SS) was the prior state-of-the art for recovering VerbNet classes. The unsupervised Joint model is competitive with Step-wise baseline, especially with the addition of `slot` features. Adding semi-supervision to the Joint model is computationally simpler and ultimately produces a superior result.

Model	**iPU** on C_1	**iPU** on C_2
Step-wise	52.21 ± 1.90	55.13 ± 1.29
Step-wise + SS	52.88 ± 1.62	55.84 ± 1.14
Joint	47.12 ± 2.56	41.76 ± 1.73
Joint + SS	58.23 ± 0.99	51.40 ± 0.98
Joint + `slot`	53.48 ± 4.15	42.10 ± 1.94
Joint + `slot` + SS	$\mathbf{69.14} \pm 0.27$	$\mathbf{57.99} \pm 1.14$

Table 2: Detail of inverse purity for partially-supervised VerbNet classes (C_1), and for never-observed VerbNet classes (C_2), for various models. We expect to recover partially-observed classes better with supervision, but we also see an improvement to recovery of classes that are outside the supervision set.

Bring-11.3. Although there is still some semantic similarity between Steal and Bring classes, they are not as closely related as Steal and Obtain. Steal-10.5 pertains to a change of possession, whereas Bring-11.3 pertains to a caused change of location. Presumably, the *take* seed sentences from Steal-10.5 attracted them primarily on the basis of the verb itself. We had no topics seeded with sentences from Bring-11.3 using any of the verbs from that class, which probably gave undue influence to *take* seed sentences from Steal-10.5.

The clusters with high purity scores group instances from a single VerbNet class, often gathering several different verbs. Topic 59, for example, has a purity score of 89.9% and clusters 4 verbs from VerbNet 45.6, the Calibratable Change of State class: *vary, rise, drop* and *dip*. These verbs all have multiple senses, but the cluster correctly groups the instances having to do with something changing along a scale (e.g., "The stock dropped 23%"). The most frequent incorrectly grouped verb in the cluster, *grow*, is a verb in the Calibratable Change of State class, but the sentences grouped here were not in the correct sense of change along a scale but of organic growth (e.g., "The farmer grows apples"). These should have been grouped with other verbs from the Grow-26.2 class.

A high purity score does not always indicate a cluster that mimics a VerbNet class, however. It occasionally shows that there was success in clustering instances of a single verb sense, but not in capturing multiple similar verbs. Topic 74 correctly grouped together 62 instances of stop in the sense of ending an activity, but only has 1 instance each of 2 other verbs: *break* and *discourage*. Although clustering only one sense of a polysemous verb is not an insignificant success, we would like the clusters to find similarities across verbs as well.

Including the supervision aids recovery of the VerbNet classes in inverse purity, as well. Topic 59, from above, was seeded with examples from VerbNet 45.6, including sentences with verbs like *gain, grow, dip* and *shift*. Out of 1042 sentences in the test set, we put 882 of them into the same topic. In the unsupervised setting, we grouped only 594 of these sentences together. We see a similar improvement for VerbNet 105, the Use class. Once seeded with examples from the verb *employ,* we now group 226 of the 362 sentences together, rather than 179.

However, the seeded topics didn't improve everything. We seeded one topic with examples from the VerbNet class Discover-84, using sentences with the verbs *discover* and *hear*. In the evaluation set, the only test examples came from the verb *find*, which is polysemous in the test classes (with examples from Discover-84 and from Get-13.5.1). After supervision, the instances of *find* were clustered together more strongly, and we increased the inverse purity of both Discover-84 and Get-13.5.1. However, in both cases these two classes were incorrectly conflated. Despite a close analysis of the SemLink *find* sentences, it is difficult to account for the placement of *find* in a topic seeded with *advise* sentences.

We also increased the inverse purity score for VerbNet class Say-37.7, which is one of the test classes omitted from the supervision. However, this increase is a result of incorrect lumping with seed examples. The verb *add* is polysemous, belonging to both Say-37.7 ("Elaine added a few words") and Mix-22.1 ("Herman added the computer to the network"), and we included Mix-22.1 examples in our supervision. This produces a cluster that is dominated by *add*, clustered with verbs like *convert* and *link*, which have much lower frequency. Because all examples of *add* end up in this cluster, more examples of Say-37.7 are clustered together after supervision, but we lump them in with Mix-22.1, resulting in a cluster that does not represent the Say-37.7 class. The unsupervised cluster with the most Say-37.7 examples has frequent verbs *say*, *tell*, *ask*, and *explain*, which clearly recovers the desired concept. A similar cluster is created in the partially-supervised clusters, but the test examples from *add* are not included in it.

Target VerbNet Class	Test Set Verbs	Unsupervised Model		Semi-Supervised Model	
		Frequent Verbs	Test Set Verbs	Frequent Verbs	Test Set Verbs
Calibratable-COS-45.6	rise (536) drop (122) move (56) vary (15) appreciate (4)	increase reduce grow exceed rise	rise (468) drop (49) vary (1) expect (24) drop (8) push (6) grow (6)	*grow* *gain* increase rise reduce *dip* *shift*	rise (512) drop (120) vary (8) drop (29) grow (17) vary (13) hit (5) rise (3) dip (3) count (2)
Use-105	use (588)	use develop support need utilize	use (369) use (5) need (3) call (2)	use need create support have *employ*	use (455) need (46) use (7) call (3) work (2)
Discover-84	find (122) *discover*, *hear*	find find out view work out advise	find (80) find (177)	find *advise* base focus depend	find(122) find (202) work (4) count (3)
Say-37.7	add (282) disclose (155) declare (46) write (26) observe (9)	say tell ask explain add	add (134) declare (9) disclose (2) admit (4) call (2)	*add* convert link subscribe append	add (282) add (3)

Table 3: Best clusters from the unsupervised and partially-supervised clustering algorithms for 4 target VerbNet classes. The most-frequent verbs in each cluster are shown, with all terms that seeded the given cluster in the semi-supervised model indicated in italics. We also show the test set verbs assigned to that cluster, with the number of sentences indicated in parentheses. Terms highlighted in gray are the model's errors, and show sentences assigned to the cluster that are not in the target VerbNet class. Verbs in both black and gray in the same cluster indicate multiple senses of the verb which should have separated into distinct clusters.

Conclusions

Our work extends efforts to produce VerbNet-like clusters from corpus data, and improves on state-of-the-art results using a computationally efficient and easy-to-implement partial supervision system. Latent variable assignments in models like LDA produce a clustering of the data, and with our technique we can efficiently guide the clustering to discover structures we care about, without requiring any change to the inference procedure after initialization. This supervision improves both the modified purity and the inverse purity of the resulting clusters, and makes transparent use of labeled data that was not possible in previous, step-wise verb clustering frameworks. Although there is still room for improvement on this task, the improvement in quality from a relatively small amount of labeled data is promising for development of VerbNet-like resources for specific domains, or in other languages, and may actually give annotators tools to improve English VerbNet.

We also demonstrate an improvement in score by double-counting our vocabulary items at two levels of abstraction, keeping `slot` and `slot:token` features side-by-side. This improvement holds independently of supervision, which highlights the importance of syntactic features for building Levin-style verb clusters. But it also suggests a natural experiment for future work, abstracting tokens into semantic categories to capture selectional preferences.

Partial supervision with direct observations, though it is a small technical change, produces a large and reliable effect on the resulting clusters. This observation is a major contribution of our work, and has applicability to all supervised and human-in-the-loop topic modeling systems. It requires considerably less computational overhead than other methods for supervised topic modeling, and produces a pronounced effect after a very limited amount of labeling.

Future Work

VerbNet has been expanded and revised in the years since the SemLink annotation was done, significantly increasing the coverage of highly frequent verbs and improving the consistency of the classes. A new version of SemLink that reflects these changes is scheduled for release in coming months. We would like to test our system with the new data, as the improved SemLink may produce a further improvement on the clusters.

In order to have impact for VerbNet annotation efforts, the model's output must be provided to annotators. We believe a tool that suggests classes, class members, and provides the annotators with a view into the example sentences would dramatically improve and accelerate their work. The tool could track annotators' decisions, allowing each annotation session to refine and further improve the model's output. The most obvious next step for this research is to build that tool, and use it to expand and improve VerbNet in as many languages as possible. In languages where no labeled exists, the unsupervised model can be used to generate candidate VerbNet clusters, which the annotators can accept, reject, or modify. These annotations can be used as supervision to generate refined clusters, improving the next batch of candidate clusters. The annotations have a strong and reliable effect on the clusters generated, so this interactive approach should allow the annotators to make rapid improvements to the clusters.

Presenting annotators with suggestions from our probabilistic verb clustering is only the first step toward a line of research in how to best supervise these probabilistic clustering models. We saw examples where supervision encouraged distinct senses to be incorrectly linked, and that this created poorer semantic clusters. We hypothesize that properly seeding the different senses into different classes will fix this issue, but this validation, and the identification of similar issues, is left to future work. Also, in this paper we have specified supervision at the sentence level, affecting counts for both topics and document distributions by labeling specific sentences, but inference-level supervision of topics can easily be applied to topics or documents without specifying particular sentences. Annotators will likely find use for both broad, topic-level supervision and focused corrections for individual sentences. Given the preliminary evidence that probabilistic clustering models respond positively and predictably to annotation, the line of research into best practices for cluster annotation seems promising.

References

Aldezabal, I.; Aranzabe, M. J.; de Ilarraza Sánchez, A. D.; and Estarrona, A. 2010. Building the basque propbank. In *LREC*.

Andrzejewski, D.; Zhu, X.; and Craven, M. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning*, 25–32. ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *the Journal of Machine Learning Research* 3:993–1022.

Bonial, C.; Stowe, K.; and Palmer, M. 2013a. Renewing and revising semlink. In *GenLex Workshop on Linked Data in Linguistics*.

Bonial, C.; Stowe, K.; and Palmer, M. 2013b. Renewing and revising semlink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, 9–17.

Brew, C., and Schulte im Walde, S. 2002. Spectral clustering for german verbs. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 117–124. Association for Computational Linguistics.

Christensen, J.; Soderland, S.; Etzioni, O.; et al. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, 52–60. Association for Computational Linguistics.

Danlos, L.; Nakamura, T.; and Pradet, Q. 2014. Toward a french verbenet (vers la création d'un verbnet français)[in french]. In *TALN-RECITAL 2014 Workshop FondamenTAL 2014: Ressources lexicales et TAL-vue d'ensemble sur les dictionnaires électroniques de Jean Dubois et Françoise Dubois-Charlier (FondamenTAL 2014: Lexical Resources and NLP)*, 103–108.

Estarrona, A.; Aldezabal, I.; Díaz de Ilarraza, A.; and Aranzabe, M. J. 2015. A methodology for the semiautomatic annotation of epec-rolsem, a basque corpus labeled at predicate level following the propbank-verbnet model. *Digital scholarship in the humanities* 31(3):470–492.

Giuglea, A.-M., and Moschitti, A. 2006. Semantic role labeling via framenet, verbnet and propbank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 929–936. Association for Computational Linguistics.

Hartmann, S.; Eckle-Kohler, J.; and Gurevych, I. 2016. Generating training data for semantic role labeling based on label transfer from linked lexical resources. *Transactions of the Association for Computational Linguistics* 4:197–213.

Hautli-Janisz, A.; King, T. H.; and Ramchand, G. 2015. Encoding event structure in urdu/hindi verbnet. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 25–33.

Hu, Y.; Boyd-Graber, J.; Satinoff, B.; and Smith, A. 2014. Interactive topic modeling. *Machine learning* 95(3):423–469.

Im Walde, S. S. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 747–753. Association for Computational Linguistics.

Jagarlamudi, J.; Daumé III, H.; and Udupa, R. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 204–213. Association for Computational Linguistics.

Kawahara, D.; Peterson, D. W.; and Palmer, M. 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)*.

Kipper, K.; Dang, H. T.; Palmer, M.; et al. 2000. Class-based construction of a verb lexicon. *AAAI/IAAI* 691:696.

Kipper, K.; Korhonen, A.; Ryant, N.; and Palmer, M. 2006.

Extending verbnet with novel verb classes. In *LREC*, 1027–1032. Citeseer.

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

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

Lacoste-Julien, S.; Sha, F.; and Jordan, M. I. 2009. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, 897–904.

Lapata, M., and Brew, C. 2004. Verb class disambiguation using informative priors. *Computational Linguistics* 30(1):45–73.

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

Li, J., and Brew, C. 2008. Which are the best features for automatic verb classification. *Proceedings of ACL-08: HLT* 434–442.

Majewska, O.; McCarthy, D.; Vulić, I.; and Korhonen, A. 2018. Acquiring verb classes through bottom-up semantic verb clustering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Marcus, M.; Kim, G.; Marcinkiewicz, M. A.; MacIntyre, R.; Bies, A.; Ferguson, M.; Katz, K.; and Schasberger, B. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, 114–119. Stroudsburg, PA, USA: Association for Computational Linguistics.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *Advances in neural information processing systems*, 121–128.

Moreda, P.; Llorens, H.; Saquete, E.; and Palomar, M. 2011. Combining semantic information in question answering systems. *Information Processing & Management* 47(6):870–885.

Mousser, J. 2010. A large coverage verb taxonomy for arabic. In *LREC*.

Parker, R.; Graff, D.; Kong, J.; Chen, K.; and Maeda, K. 2011. English gigaword fifth edition ldc2011t07.

Peterson, D. W., and Palmer, M. 2018. Bayesian verb sense clustering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 5398–5405.

Peterson, D. W.; Boyd-Graber, J.; Palmer, M.; and Kawhara, D. 2016. Leveraging verbnet to build corpus-specific verb clusters. ** SEM* 102.

Pradet, Q.; Danlos, L.; and De Chalendar, G. 2014. Adapting verbnet to french using existing resources. In *LREC'14-Ninth International Conference on Language Resources and Evaluation*.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 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*, 248–256. Association for Computational Linguistics.

Scarton, C.; Duran, M. S.; and Aluísio, S. M. 2014. Using cross-linguistic knowledge to build verbnet-style lexicons: Results for a (brazilian) portuguese verbnet. In *International Conference on Computational Processing of the Portuguese Language*, 149–160. Springer.

Shen, D., and Lapata, M. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

Sun, L., and Korhonen, A. 2011. Hierarchical verb clustering using graph factorization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1023–1033. Association for Computational Linguistics.

Vulić, I.; Mrkšić, N.; and Korhonen, A. 2017. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2546–2558.

Xie, P.; Yang, D.; and Xing, E. 2015. Incorporating word correlation knowledge into topic modeling. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: human language technologies*, 725–734.

Yang, Y.; Downey, D.; and Boyd-Graber, J. 2015. Efficient methods for incorporating knowledge into topic models. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 308–317.