A Step-wise Usage-based Method for Inducing Polysemy-aware Verb Classes

Daisuke Kawahara† **Daniel W. Peterson**‡ **Martha Palmer**‡
†Kyoto University, Kyoto, Japan
‡University of Colorado at Boulder, Boulder, CO, USA
dk@i.kyoto-u.ac.jp, {Daniel.W.Peterson, Martha.Palmer}@colorado.edu

Abstract

We present an unsupervised method for inducing verb classes from verb uses in giga-word corpora. Our method consists of two clustering steps: verb-specific semantic frames are first induced by clustering verb uses in a corpus and then verb classes are induced by clustering these frames. By taking this step-wise approach, we can not only generate verb classes based on a massive amount of verb uses in a scalable manner, but also deal with verb polysemy, which is bypassed by most of the previous studies on verb clustering. In our experiments, we acquire semantic frames and verb classes from two giga-word corpora, the larger comprising 20 billion words. The effectiveness of our approach is verified through quantitative evaluations based on polysemy-aware gold-standard data.

1 Introduction

A verb plays a primary role in conveying the meaning of a sentence. Capturing the sense of a verb is essential for natural language processing (NLP), and thus lexical resources for verbs play an important role in NLP.

Verb classes are one such lexical resource. Manually-crafted verb classes have been developed, such as Levin's classes (Levin, 1993) and their extension, VerbNet (Kipper-Schuler, 2005), in which verbs are organized into classes on the basis of their syntactic and semantic behavior. Such verb classes have been used in many NLP applications that need to consider semantics in particular, such as word sense disambiguation (Dang, 2004), semantic parsing (Swier and Stevenson, 2005; Shi and Mihalcea, 2005) and discourse parsing (Subba and Di Eugenio, 2009).

There have also been many attempts to automatically acquire verb classes with the goal of either adding frequency information to an existing resource or of inducing similar verb classes for other languages. Most of these approaches assume that all target verbs are monosemous (Stevenson and Joanis, 2003; Schulte im Walde, 2006; Joanis et al., 2008; Li and Brew, 2008; Sun et al., 2008; Sun and Korhonen, 2009; Vlachos et al., 2009; Parisien and Stevenson, 2010; Parisien and Stevenson, 2011; Falk et al., 2012; Lippincott et al., 2012; Reichart and Korhonen, 2013; Sun et al., 2013). This monosemous assumption, however, is not realistic because many frequent verbs actually have multiple senses. Moreover, to the best of our knowledge, none of the following approaches attempt to quantitatively evaluate soft clusterings of verb classes induced by polysemy-aware unsupervised approaches (Korhonen et al., 2003; Lapata and Brew, 2004; Li and Brew, 2007; Schulte im Walde et al., 2008).

In this paper, we propose an unsupervised method for inducing verb classes that is aware of verb polysemy. Our method consists of two clustering steps: verb-specific semantic frames are first induced by clustering verb uses in a corpus and then verb classes are induced by clustering these frames. By taking this step-wise approach, we can not only induce verb classes with frequency information from a massive amount of verb uses in a scalable manner, but also deal with verb polysemy.

Our novel contributions are summarized as follows:

- induce both semantic frames and verb classes from a massive amount of verb uses by a scalable method,
- explicitly deal with verb polysemy,
- discover effective features for each of the clustering steps, and
- quantitatively evaluate a soft clustering of verbs.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1030–1040,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

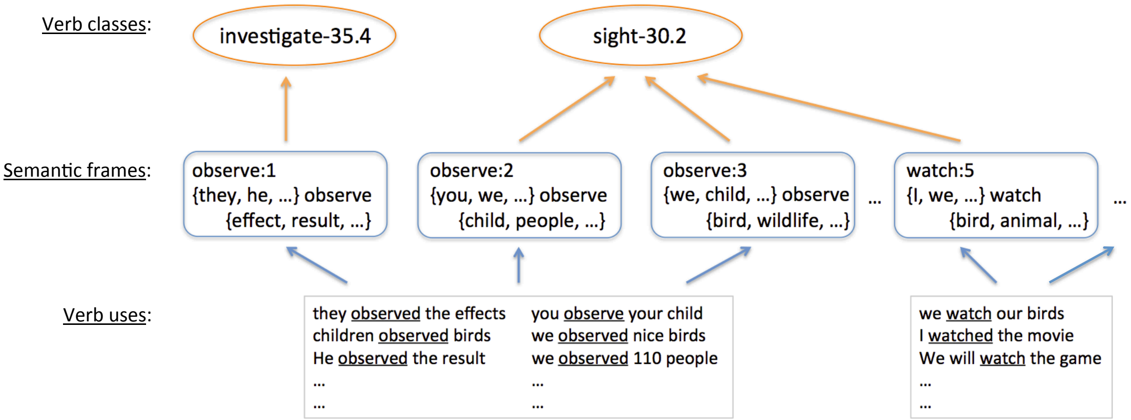

Figure 1: Overview of our two-step approach. Verb-specific semantic frames are first induced from verb uses (lower part) and then verb classes are induced from the semantic frames (upper part). The labels of verb classes are manually assigned here for better understanding.

2 Related Work

As stated in Section 1, most of the previous studies on verb clustering assume that verbs are monosemous. A typical method in these studies is to represent each verb as a single data point and apply classification (e.g., Joanis et al. (2008)) or clustering (e.g., Sun and Korhonen (2009)) to these data points. As a representation for a data point, distributions of subcategorization frames are often used, and other semantic features (e.g., selectional preferences) are sometimes added to improve the performance.

Among these studies on monosemous verb clustering (i.e., predominant class induction), there have been several Bayesian methods. Vlachos et al. (2009) proposed a Dirichlet process mixture model (DPMM; Neal (2000)) to cluster verbs based on subcategorization frame distributions. They evaluated their result with a gold-standard test set, where a single class is assigned to a verb. Parisien and Stevenson (2010) proposed a hierarchical Dirichlet process (HDP; Teh et al. (2006)) model to jointly learn argument structures (subcategorization frames) and verb classes by using syntactic features. Parisien and Stevenson (2011) extended their model by adding semantic features. They tried to account for verb learning by children and did not evaluate the resultant verb classes. Modi et al. (2012) extended the model of Titov and Klementiev (2012), which is an unsupervised model for inducing semantic roles, to jointly induce semantic roles and frames across verbs using the Chinese Restaurant Process (Aldous, 1985). All of the above methods considered verbs to be monosemous and did not deal with verb polysemy. Our approach also uses Bayesian methods, but is designed to capture verb polysemy.

We summarize a few studies that consider polysemy of verbs in the rest of this section.

Miyao and Tsujii (2009) proposed a supervised method that can handle verb polysemy. Their method represents a verb's syntactic and semantic features, and learns a log-linear model from the SemLink corpus (Loper et al., 2007). Boleda et al. (2007) also proposed a supervised method for Catalan adjectives considering the polysemy of adjectives.

The most closely related work to our polysemy-aware task of unsupervised verb class induction is the work of Korhonen et al. (2003), who used distributions of subcategorization frames to cluster verbs. They adopted the Nearest Neighbor (NN) and Information Bottleneck (IB) methods for clustering. In particular, they tried to consider verb polysemy by using the IB method, which is a soft clustering method (Tishby et al., 1999). However, the verb itself is still represented as a single data point. After performing soft clustering, they noted that most verbs fell into a single class, and they decided to assign a single class to each verb by hardening the clustering. They considered multiple classes only in the gold-standard data used for their evaluations. We also evaluate our induced verb classes on this gold-standard data, which was created on the basis of Levin's classes (Levin, 1993).

Lapata and Brew (2004) and Li and Brew (2007) proposed probabilistic models for calculating prior probabilities of verb classes for a verb. These models are approximated to condition not on verbs but on subcategorization frames. As mentioned in Li and Brew (2007), it is desirable to extend the model to depend on verbs to further improve accuracy. They conducted several evaluations including predominant class induction and token-level verb sense disambiguation, but did not evaluate multiple classes output by their models. Schulte im Walde et al. (2008) also applied probabilistic soft clustering to verbs by incorporating subcategorization frames and selectional preferences based on WordNet. This model is based on the Expectation-Maximization algorithm and the Minimum Description Length principle. Since they focused on the incorporation of selectional preferences, they did not evaluate verb classes but evaluated only selectional preferences using a language model-based measure.

Materna proposed LDA-frames, which are defined across verbs and can be considered to be a kind of verb class (Materna, 2012; Materna, 2013). LDA-frames are probabilistic semantic frames automatically induced from a raw corpus. He used a model based on latent Dirichlet allocation (LDA; Blei et al. (2003)) and the Dirichlet process to cluster verb instances of a triple (subject, verb, object) to produce semantic frames and roles. Both of these are represented as a probabilistic distribution of words across verbs. He applied this method to the BNC and acquired 1,200 frames and 400 roles (Materna, 2012). He did not evaluate the resulting frames as verb classes.

In sum, there have been no studies that quantitatively evaluate polysemous verb classes automatically induced by unsupervised methods.

3 Our Approach

3.1 Overview

Our objective is to automatically learn semantic frames and verb classes from a massive amount of verb uses following usage-based approaches. Although Bayesian approaches are a possible solution to simultaneously induce frames and verb classes from a corpus as used in previous studies, it has prohibitive computational cost. For instance, Parisien and Stevenson applied HDP only to a small-scale child speech corpus that contains 170K verb uses to jointly induce subcategorization frames and verb classes (Parisien and Stevenson, 2010; Parisien and Stevenson, 2011). Materna applied an LDA-based method to the BNC, which contains 1.4M verb uses, to induce semantic frames across verbs that can be considered to be verb classes (Materna, 2012; Materna, 2013). However, it would take three months for this experiment using this 100 million word corpus.[1] Although it is best to use the largest possible corpus for this kind of knowledge acquisition tasks (Sasano et al., 2009), it is infeasible to scale to giga-word corpora using such joint models.

In this paper, we propose a two-step approach for inducing semantic frames and verb classes. First, we make multiple data points for each verb to deal with verb polysemy (cf. polysemy-aware previous studies still represented a verb as one data point (Korhonen et al., 2003; Miyao and Tsujii, 2009)). To do that, we induce verb-specific semantic frames by clustering verb uses. Then, we induce verb classes by clustering these verb-specific semantic frames across verbs. An interesting point here is that we can use exactly the same method for these two clustering steps.

Our procedure to automatically induce verb classes from verb uses is summarized as follows:

1. induce verb-specific semantic frames by clustering predicate-argument structures for each verb extracted from automatic parses as shown in the lower part of Figure 1, and
2. induce verb classes by clustering the induced semantic frames across verbs as shown in the upper part of Figure 1.

Each of these two steps is described in the following sections in detail.

3.2 Inducing Verb-specific Semantic Frames

We induce verb-specific semantic frames from verb uses based on the method of Kawahara et al. (2014). Our semantic frames consist of case slots, each of which consists of word instances that can be filled. The procedure for inducing these semantic frames is as follows:

1. apply dependency parsing to a raw corpus and extract predicate-argument structures for each verb from the automatic parses,
2. merge the predicate-argument structures that have presumably the same meaning based on the assumption of one sense per collocation (Yarowsky, 1993) to get a set of initial frames, and

[1] In our replication experiment, it took a week to perform 70 iterations using Materna's code and an Intel Xeon E5-2680 (2.7GHz) CPU. To reach 1,000 iterations, which are reported to be optimum, it would take three months.

3. apply clustering to the initial frames based on the Chinese Restaurant Process (Aldous, 1985) to produce verb-specific semantic frames.

These three steps are briefly described below.

3.2.1 Extracting Predicate-argument Structures from a Raw Corpus

We apply dependency parsing to a large raw corpus. We use the Stanford parser with Stanford dependencies (de Marneffe et al., 2006).[2] Collapsed dependencies are adopted to directly extract prepositional phrases.

Then, we extract predicate-argument structures from the dependency parses. Dependents that have the following dependency relations to a verb are extracted as arguments:

> nsubj, xsubj, dobj, iobj, ccomp, xcomp, prep_*

In this process, the verb and arguments are lemmatized, and only the head of an argument is preserved for compound nouns.

Predicate-argument structures are collected for each verb and the subsequent processes are applied to the predicate-argument structures of each verb.

3.2.2 Constructing Initial Frames from Predicate-argument Structures

To make the computation feasible, we merge the predicate-argument structures that have the same or similar meaning to get initial frames. These initial frames are the input of the subsequent clustering process. For this merge, we assume one sense per collocation (Yarowsky, 1993) for predicate-argument structures.

For each predicate-argument structure of a verb, we couple the verb and an argument to make a unit for sense disambiguation. We select an argument in the following order by considering the degree of effect on the verb sense:[3]

> dobj, ccomp, nsubj, prep_*, iobj.

Then, the predicate-argument structures that have the same verb and argument pair (slot and word, e.g., “dobj:effect”) are merged into an initial frame. After this process, we discard minor initial frames that occur fewer than 10 times.

3.2.3 Clustering Method

We cluster initial frames for each verb to produce semantic frames using the Chinese Restaurant Process (Aldous, 1985), regarding each initial frame as an instance.

We calculate the posterior probability of a cluster c_j given an initial frame f_i as follows:

$$P(c_j|f_i) \propto \begin{cases} \frac{n(c_j)}{N+\alpha} \cdot P(f_i|c_j) & c_j \neq new \\ \frac{\alpha}{N+\alpha} \cdot P(f_i|c_j) & c_j = new, \end{cases} \quad (1)$$

where N is the number of initial frames for the target verb and $n(c_j)$ is the current number of initial frames assigned to the cluster c_j. α is a hyper-parameter that determines how likely it is for a new cluster to be created. In this equation, the first term is the Dirichlet process prior and the second term is the likelihood of f_i.

$P(f_i|c_j)$ is defined based on the Dirichlet-Multinomial distribution as follows:

$$P(f_i|c_j) = \prod_{w \in V} P(w|c_j)^{count(f_i,w)}, \quad (2)$$

where V is the vocabulary in all case slots cooccurring with the verb and $count(f_i, w)$ is the number of w in the initial frame f_i. The original method in Kawahara et al. (2014) defined w as pairs of slots and words, e.g., “nsubj:child” and “dobj:bird,” but does not consider slot-only features, e.g., “nsubj” and “dobj,” which ignore lexical information. Here we experiment with both representations and compare the results.

$P(w|c_j)$ is defined as follows:

$$P(w|c_j) = \frac{count(c_j, w) + \beta}{\sum_{t \in V} count(c_j, t) + |V| \cdot \beta}, \quad (3)$$

where $count(c_j, w)$ is the current number of w in the cluster c_j, and β is a hyper-parameter of Dirichlet distribution. For a new cluster, this probability is uniform ($1/|V|$).

We regard each output cluster as a semantic frame, by merging the initial frames in a cluster into a semantic frame. In this way, semantic frames for each verb are acquired.

We use Gibbs sampling to realize this clustering.

3.3 Inducing Verb Classes from Semantic Frames

To induce verb classes across verbs, we apply clustering to the induced verb-specific semantic

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] If a predicate-argument structure has multiple prepositional phrases, one of them is randomly selected.

frames. We can use exactly the same clustering method as described in Section 3.2.3 by using semantic frames for multiple verbs as an input instead of initial frames for a single verb. This is because an initial frame has the same structure as a semantic frame, which is produced by merging initial frames. We regard each output cluster as a verb class this time.

For the features, w, in equation (2), we try the two representations again: slot-only features and slot-word pair features. The representation using only slots corresponds to the consideration of only syntactic argument patterns. The other representation using the slot-word pairs means that semantic similarity based on word overlap is naturally considered by looking at lexical information. We will compare in our experiments four possible combinations: two feature representations for each of the two clustering steps.

4 Experiments and Evaluations

We first describe our experimental settings and define evaluation metrics to evaluate induced soft clusterings of verb classes. Then, we conduct type-level multi-class evaluations, type-level single-class evaluations and token-level multi-class evaluations. These two levels of evaluations are performed by considering the work of Reichart et al. (2010) on clustering evaluation. Finally, we discuss the results of our full experiments.

4.1 Experimental Settings

We use two kinds of large-scale corpora: a web corpus and the English Gigaword corpus.

To prepare a web corpus, we extracted sentences from crawled web pages that are judged to be written in English based on the encoding information. Then, we selected sentences that consist of at most 40 words, and removed duplicated sentences. From this process, we obtained a corpus of one billion sentences, totaling approximately 20 billion words. We focused on verbs whose frequency in the web corpus was more than 1,000. There were 19,649 verbs, including phrasal verbs, and separating passive and active constructions. We extracted 2,032,774,982 predicate-argument structures.

We also used the English Gigaword corpus (LDC2011T07; English Gigaword Fifth Edition). This corpus consists of approximately 180 million sentences, which totaling four billion words. There were 7,356 verbs after applying the same frequency threshold as the web corpus. We extracted 423,778,278 predicate-argument structures from this corpus.

We set the hyper-parameters α in (1) and β in (3) to 1.0. The cluster assignments for all the components were initialized randomly. We took 100 samples for each input frame and selected the cluster assignment that has the highest probability.

4.2 Evaluation Metrics

To measure the precision and recall of a clustering, modified purity and inverse purity (also called collocation or weighted class accuracy) are commonly used in previous studies on verb clustering (e.g., Sun and Korhonen (2009)). However, since these measures are only applicable to a hard clustering, it is necessary to extend them to be applicable to a soft clustering, because in our task a verb can belong to multiple clusters or classes.[4] We propose a normalized version of modified purity and inverse purity. This kind of normalization for soft clusterings was performed for other evaluation metrics as in Springorum et al. (2013).

To measure the precision of a clustering, a normalized version of modified purity is defined as follows. Suppose K is the set of automatically induced clusters and G is the set of gold classes. Let K_i be the verb vector of the i-th cluster and G_j be the verb vector of the j-th gold class. Each component of these vectors is a normalized frequency, which equals a cluster/class attribute probability given a verb. Where there is no frequency information available for class distribution, such as the gold-standard data described in Section 4.3, we use a uniform distribution across the verb's classes. The core idea of purity is that each cluster K_i is associated with its most prevalent gold class. In addition, to penalize clusters that consist of only one verb, such singleton clusters in K are considered as errors, as is usual with modified purity. The normalized modified purity (nmPU) can then be written as follows:

$$\text{nmPU} = \frac{1}{N} \sum_{i \text{ s.t. } |K_i|>1} \max_j \delta_{K_i}(K_i \cap G_j), \quad (4)$$

$$\delta_{K_i}(K_i \cap G_j) = \sum_{v \in K_i \cap G_j} c_{iv}, \quad (5)$$

[4] Korhonen et al. (2003) evaluated hard clusterings based on a gold standard with multiple classes per verb. They reported only precision measures including modified purity, and avoided extending the evaluation metrics for soft clusterings.

verb	classes	verb	classes
place	**9**	drop	**9**, 45, 004, 47,
dye	**24**, 21, 41		51, A54, A30
focus	**31**, 45	bake	**26**, 45
stare	**30**	persuade	**002**
lay	**9**	sparkle	**43**
build	**26**, 45	pour	**9**, 43, 26, 57,
force	**002**, 11		13, 31
glow	**43**	invent	**26**, 27

Table 1: An excerpt of the gold-standard verb classes for several verbs from Korhonen et al. (2003). The classes starting with '0' were derived from the LCS database, those starting with 'A' were defined by Korhonen et al., and the other classes were from Levin's classes. A bolded class is the predominant class for each verb.

where N denotes the total number of verbs, $|K_i|$ denotes the number of positive components in K_i, and c_{iv} denotes the v-th component of K_i. $\delta_{K_i}(K_i \cap G_j)$ means the total mass of the set of verbs in $K_i \cap G_j$, given by summing up the values in K_i. In case of evaluating a hard clustering, this is equal to $|K_i \cap G_j|$ because all the values of c_{iv} are equal to 1.

As usual, the following normalized inverse purity (niPU) is used to measure the recall of a clustering:

$$\text{niPU} = \frac{1}{N} \sum_j \max_i \delta_{G_j}(K_i \cap G_j). \qquad (6)$$

Finally, we use the harmonic mean (F$_1$) of nmPU and niPU as a single measure of clustering quality.

4.3 Type-level Multi-class Evaluations

We first evaluate our induced verb classes on the test set created by Korhonen et al. (2003) (Table 1 of their paper) which was created by considering verb polysemy on the basis of Levin's classes and the LCS database (Dorr, 1997). It consists of 62 classes and 110 verbs, out of which 35 verbs are monosemous and 75 verbs are polysemous. The average number of verb classes per verb is 2.24. An excerpt from this data is shown in Table 1.

As our baselines, we adopt two previously proposed methods. We first implemented a soft clustering method for verb class induction proposed by Korhonen et al. (2003). They used the information bottleneck (IB) method for assigning probabilities of classes to each verb. Note that Korhonen et al. (2003) actually hardened the clusterings and left the evaluations of soft clusterings for their future work. For input data, we employ VALEX (Korhonen et al., 2006), which is a publicly-available large-scale subcategorization lexicon.[5] By following the method of Korhonen et al. (2003), prepositional phrases (pp) are parameterized for two frequent subcategorization frames (NP and NP_PP), and the unfiltered raw frequencies of subcategorization frames are used as features to represent a verb. It is necessary to specify the number of clusters, k, for the IB method beforehand, and we adopt 35 and 42 clusters according to their reported high accuracies. To output multiple classes for each verb, we set a threshold, t, for class attribute probabilities. That is, classes that have a higher class attribute probability than the threshold are output for each verb. We report the results of the following threshold values: 0.01, 0.02, 0.05 and 0.10.

method	K	nmPU	niPU	F$_1$
IB (k=35, t=0.10)	35.0	53.59	51.44	52.44
IB (k=35, t=0.05)	35.0	53.67	52.62	53.10
IB (k=35, t=0.02)	35.0	54.42	54.43	54.40
IB (k=35, t=0.01)	35.0	54.60	55.54	55.04
IB (k=42, t=0.10)	41.6	55.42	49.46	52.24
IB (k=42, t=0.05)	41.8	55.55	49.97	52.59
IB (k=42, t=0.02)	42.0	56.19	51.24	53.58
IB (k=42, t=0.01)	42.0	56.80	51.92	54.24
LDA-frames (t=0.10)	100	47.52	56.83	51.76
LDA-frames (t=0.05)	165	50.46	67.94	57.91
LDA-frames (t=0.02)	306	49.98	75.50	60.14
LDA-frames (t=0.01)	458	49.55	82.71	61.97
Gigaword/S-S	272.8	63.46	67.66	65.49
Gigaword/S-SW	36.4	31.49	95.70	47.38
Gigaword/SW-S	186.2	63.52	64.18	63.84
Gigaword/SW-SW	30.0	36.27	94.66	52.40
web/S-S	363.6	61.32	78.64	68.90
web/S-SW	52.2	35.80	**99.30**	52.62
web/SW-S	212.2	**66.26**	77.38	**71.39**
web/SW-SW	55.0	36.70	96.25	53.13

Table 2: Type-level multi-class evaluations. K represents the (average) number of induced classes. "S" denotes the use of slot-only features and "SW" denotes the use of slot-word pair features. For example, "SW-S" means that slot-word pair features are used for semantic frame induction and slot-only features are used for verb class induction.

The other baseline is LDA-frames (Materna, 2012). We use the induced LDA-frames that are

[5] http://ilexir.co.uk/applications/valex/

		predominant class eval			multiple class eval		
method	K	mPU	iPU	F_1	mPU	niPU	F_1
NN	24	46.36	52.73	49.34	52.73	46.85	49.62
IB (*k*=35)	34.8	42.73	51.82	46.82	51.64	46.83	49.09
IB (*k*=42)	41.0	**47.45**	50.91	49.11	**55.27**	45.45	49.87
LDA-frames	53	30.00	47.27	36.71	41.82	44.28	43.01
Gigaword/S	9.6	25.64	71.27	37.70	32.91	64.71	43.62
Gigaword/SW	10.6	30.36	71.09	42.25	39.82	66.92	49.70
web/S	20.4	42.73	61.46	**50.31**	54.91	57.12	55.86
web/SW	11.8	34.36	**71.82**	46.40	49.09	**67.01**	**56.50**

Table 3: Type-level single-class evaluations against predominant/multiple classes. K represents the (average) number of induced classes.

available on the web site.[6] This frame data was induced from the BNC and consists of 1,200 frames and 400 semantic roles. Again, we set a threshold for frame attribute probabilities.

We report results using our methods with four feature combinations (slot-only (S) and slot-word pair (SW) features each used for both the frame-generation and verb-class clustering steps) for both the Gigaword and web corpora. Table 2 lists evaluation results for the baseline methods and our methods.[7] The results of the IB baseline and our methods are obtained by averaging five runs.

We can see that "web/SW-S" achieved the best performance and obtained a higher F_1 than the baselines by more than nine points. "Web/SW-S" uses the combination of slot-word pair features for clustering verb-specific frames and slot-only features for clustering across verbs. Interestingly, this result indicates that slot distributions are more effective than lexical information in slot-word pairs for inducing verb classes similar to the gold standard. This result is consistent with expectations, given a gold standard based on Levin's verb classes, which are organized according to the syntactic behavior of verbs. The use of slot-word pairs for verb class induction generally merged too many frames into each class, apparently due to accidental word overlaps across verbs.

The verb classes induced from the web corpus achieved a higher F_1 than those from the Gigaword corpus. This can be attributed to the larger size of the web corpus. The employment of this kind of huge corpus is enabled by our scalable method.

4.4 Type-level Single-class Evaluations against Predominant/Multiple Classes

Since we focus on the handling of verb polysemy, predominant class induction for each verb is not our main objective. However, we wish to compare our method with previous work on the induction of a predominant (monosemous) class for each verb.

To output a single class for each verb by using our proposed method, we skip the induction of verb-specific semantic frames and instead create a single frame for each verb by merging all predicate-argument structures of the verb. Then, we apply clustering to these frames across verbs. For clustering features, we again compare two representations: slot-only features (S) and slot-word pair features (SW).

We evaluate the single-class output for each verb based on the predominant gold-standard classes, which are defined for each verb in the test set of Korhonen et al. (2003). This data contains 110 verbs and 33 classes. We evaluate these single-class outputs in the same manner as Korhonen et al. (2003), using the gold standard with multiple classes, which we also use for our multi-class evaluations.

As we did with the multi-class evaluations, we adopt modified purity (mPU), inverse purity (iPU) and their harmonic mean (F_1) as the metrics for the evaluation with predominant classes. It is not necessary to normalize these metrics when we treat verbs as monosemous, and evaluate against the predominant sense. When we evaluate against the multiple classes in the gold standard, we do normalize the inverse purity.

For baselines, we once more adopt the Nearest Neighbor (NN) and Information Bottleneck (IB) methods proposed by Korhonen et al. (2003), and LDA-frames proposed by Materna (2012). The

[6] http://nlp.fi.muni.cz/projekty/lda-frames/

[7] Although we do not think that the classes with very small attribute probabilities are meaningful, the F_1 scores for lower thresholds than 0.01 converged to about 66 in the case of LDA-frames.

clusterings with the NN and IB methods are obtained by using the VALEX subcategorization lexicon. To harden the clusterings of the IB method and the LDA-frames, the class with the highest probability is selected for each verb. This hardening process is exactly the same as Korhonen et al. (2003). Note that our results of the NN and IB methods are different from those reported in their paper since the data source is different.[8]

Table 3 lists accuracies of baseline methods and our methods. Our proposed method using the web corpus achieved comparable performance with the baseline methods on the predominant class evaluation and outperformed them on the multiple class evaluation. More sophisticated methods for predominant class induction, such as the method of Sun and Korhonen (2009) using selectional preferences, could produce better single-class outputs, but have difficulty in producing polysemy-aware verb classes.

From the result, we can see that the induced verb classes based on slot-only features did not achieve a higher F_1 than those based on slot-word pair features in many cases. This result is different from that of multi-class evaluations in Section 4.3. We speculate that slot distributions are not so different among verbs when all uses of a verb are merged into one frame, and thus their discrimination power is lower than that in the intermediate construction of semantic frames.

4.5 Token-level Multi-class Evaluations

We conduct token-level multi-class evaluations using 119 verbs, which appear 100 or more times in sections 02-21 of the SemLink WSJ corpus. These 119 verbs cover 102 VerbNet classes, and 48 of them are polysemous in the sense of being in more than one VerbNet class. Each instance of these 119 verbs in this corpus belongs to one of 102 VerbNet classes. We first add these instances to the instances from a raw corpus and apply the two-step clustering to these merged instances. Then, we compare the induced verb classes of the SemLink instances with their gold-standard VerbNet classes. We report the values of modified purity (mPU), inverse purity (iPU) and their harmonic mean (F_1). It is not necessary to normalize these metrics because the clustering of these instances is hard.

method	K	mPU	iPU	F_1
Gigaword/S-NIL	–	93.43	20.06	33.03
Gigaword/SW-NIL	–	94.45	41.07	57.25
Gigaword/S-S	512.2	75.06	45.26	56.47
Gigaword/SW-S	260.6	73.98	56.45	64.04
web/S-NIL	–	93.70	32.96	48.76
web/SW-NIL	–	**94.51**	44.95	60.92
web/S-S	500.0	72.25	52.48	60.79
web/SW-S	255.2	72.65	**61.00**	**66.31**

Table 4: Token-level evaluations against VerbNet classes. K represents the average number of induced classes.

For clustering features, we compare two feature combinations: “S-S” and “SW-S,” which achieved high performance in the type-level multi-class evaluations (Section 4.3). The results of these methods are obtained by averaging five runs. For a baseline, we use verb-specific semantic frames without clustering across verbs (“S-NIL” and “SW-NIL”), where these frames are considered to be verb classes but not shared across verbs. Table 4 lists accuracies of these methods for the two corpora. We can see that “SW-S” achieved a higher F_1 than “S-S” and the baselines without verb class induction (“S-NIL” and “SW-NIL”).

Modi et al. (2012) induced semantic frames across verbs using the monosemous assumption and reported an F_1 of 44.7% (77.9% PU and 31.4% iPU) for the assignment of FrameNet frames to the FrameNet corpus. We also conducted the above evaluation against FrameNet frames for 75 verbs.[9] We achieved an F_1 of 62.79% (66.97% mPU and 59.09% iPU) for “web/SW-S,” and an F_1 of 60.06% (65.58% mPU and 55.39% iPU) for “Gigaword/SW-S.” It is difficult to directly compare these results with Modi et al. (2012), but our induced verb classes seem to have higher F_1 accuracy.

4.6 Full Experiments and Discussions

We finally induce verb classes from the semantic frames of 1,667 verbs, which appear at least once in sections 02-21 of the WSJ corpus. Based on the best results in the above evaluations, we induced semantic frames using slot-word pair features, and then induced verb classes using slot-only features. We ended with 38,481 semantic frames and 699 verb classes from the Gigaword

[8] Korhonen et al. (2003) reported that the highest modified purity was 49% against predominant classes and 60% against multiple classes.

[9] Since FrameNet frames are not assigned to all verbs of SemLink, the number of verbs is different from the evaluations against VerbNet classes.

class	semantic frames
Class 1	rave:1, talk:1
Class 2	need:2, say:2
Class 3	smell:1, sound:1
Class 4	concentrate:1, focus:1
Class 5	express:2, inquire:62, voice:1
Class 6	revolve:1, snake:2, wrap:2
Class 7	hand:1, hand:3, hand:4
Class 8	depend:1, rely:1, rely:3
Class 9	collaborate:1, compete:2, work:1
Class 10	coach:3, teach:3, teach:4
Class 11	dance:1, react:1, stick:1
Class 12	advise:8, express:4, quiz:10, voice:2
Class 13	give:18, grant:6, offer:11, offer:12
Class 14	keep:14, keep:18, stay:4, stay:488
Class 15	cuff:5, fasten:2, tie:1, tie:4
Class 16	arrange:3, book:4, make:27, reserve:5
Class 17	deport:6, differ:1, fluctuate:1, vary:1
Class 18	peek:1, peek:3, peer:1, peer:7, ...
Class 19	groan:1, growl:1, hiss:1, moan:1, purr:1
Class 20	inform:1, notify:2, remind:1, beware:1, ...

Table 5: Examples of induced verb classes. Underlined semantic frames are shown in Table 6.

corpus, and 61,903 semantic frames and 840 verb classes from the web corpus. It took two days to induce verb classes from the Gigaword corpus and three days from the web corpus.

Examples of verb classes and semantic frames induced from the web corpus are shown in Table 5 and Table 6. While there are many classes with consistent meanings, such as "Class 4" and "Class 16," some classes have mixed meanings. For instance, "Class 2" consists of the semantic frames "need:2" and "say:2." These frames were merged due to the high syntactic similarity of constituting slot distributions, which are comprised of a subject and a sentential complement. To improve the quality of verb classes, it is necessary to develop a clustering model that can consider syntactic and lexical similarity in a balanced way.

5 Conclusion

We presented a step-wise unsupervised method for inducing verb classes from instances in gigaword corpora. This method first clusters predicate-argument structures to induce verb-specific semantic frames and then clusters these semantic frames across verbs to induce verb classes. Both clustering steps are performed with exactly the same method, which is based on the Chinese Restaurant Process. The resulting semantic frames and verb classes are open to the public and also can be searched via our web interface.[10]

	slot	instance words
need:2	nsubj ccomp	you:2150273, i:7678, we:4599, ... ⟨s⟩:2193321
say:2	nsubj ccomp	she:1705781, he:20693, i:9422, ... ⟨s⟩:1829616
inform:1	nsubj dobj prep_of ⋮	i:11100, he:10323, we:6373, ... me:30646, you:27678, us:21642, ... decision:846, this:759, situation:688, ...
notify:2	nsubj dobj prep_of ⋮	we:7505, you:3439, i:1035, ... you:18604, us:7281, them:3649, ... change:1540, problem:496, status:386, ...

Table 6: Examples of induced semantic frames. The number following an instance word denotes its frequency and ⟨s⟩ denotes a sentential complement.

From the results, we can see that the combination of the slot-word pair features for clustering verb-specific frames and the slot-only features for clustering across verbs is the most effective and outperforms the baselines by approximately 10 points. This indicates that slot distributions are more effective than lexical information in slot-word pairs for the induction of verb classes, when Levin-style classes are used for evaluation. This is consistent with Levin's principle of organizing verb classes according to the syntactic behavior of verbs.

As applications of the resulting semantic frames and verb classes, we plan to integrate them into syntactic parsing, semantic role labeling and verb sense disambiguation. For instance, Kawahara and Kurohashi (2006) improved accuracy of dependency parsing based on Japanese semantic frames automatically induced from a raw corpus. It is also valuable and promising to apply the induced verb classes to NLP applications as used in metaphor identification (Shutova et al., 2010) and argumentative zoning (Guo et al., 2011).

Acknowledgments

This work was supported by Kyoto University John Mung Program and JST CREST. We also gratefully acknowledge the support of the National Science Foundation Grant NSF-IIS-1116782, A Bayesian Approach to Dynamic Lexical Resources for Flexible Language Processing. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[10] http://nlp.ist.i.kyoto-u.ac.jp/member/kawahara/cf/crp.en/

References

David Aldous. 1985. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII —1983*, pages 1–198.

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

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2007. Modelling polysemy in adjective classes by multi-label classification. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 171–180.

Hoa Trang Dang. 2004. *Investigations into the role of lexical semantics in word sense disambiguation*. Ph.D. thesis, University of Pennsylvania.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 449–454.

Bonnie J. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322.

Ingrid Falk, Claire Gardent, and Jean-Charles Lamirel. 2012. Classifying French verbs using French and English lexical resources. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 854–863.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 176–183.

Daisuke Kawahara, Daniel W. Peterson, Octavian Popescu, and Martha Palmer. 2014. Inducing example-based semantic frames from a massive amount of verb uses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

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*, pages 64–71.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 345–352.

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

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

Jianguo Li and Chris Brew. 2007. Disambiguating Levin verbs using untagged data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

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

Thomas Lippincott, Anna Korhonen, and Diarmuid Ó Séaghdha. 2012. Learning syntactic verb frames using graphical models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 420–429.

Edward Loper, Szu-Ting Yi, and Martha Palmer. 2007. Combining lexical resources: mapping between PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Linguistics*.

Jiří Materna. 2012. LDA-frames: An unsupervised approach to generating semantic frames. In *Proceedings of the 13th International Conference CICLing 2012, Part I*, pages 376–387.

Jiří Materna. 2013. Parameter estimation for LDA-frames. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 482–486.

Yusuke Miyao and Jun'ichi Tsujii. 2009. Supervised learning of a probabilistic lexicon of verb semantic classes. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1337.

Ashutosh Modi, Ivan Titov, and Alexandre Klementiev. 2012. Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 1–7.

Radford M. Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Christopher Parisien and Suzanne Stevenson. 2010. Learning verb alternations in a usage-based Bayesian model. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

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*.

Roi Reichart and Anna Korhonen. 2013. Improved lexical acquisition through DPP-based verb clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 862–872.

Roi Reichart, Omri Abend, and Ari Rappoport. 2010. Type level clustering evaluation: New measures and a POS induction case study. In *Proceedings of the 14th Conference on Computational Natural Language Learning*, pages 77–87.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2009. The effect of corpus size on case frame acquisition for discourse analysis. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 521–529.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. 2008. Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of ACL-08: HLT*, pages 496–504.

Sabine Schulte im Walde. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing*, pages 100–111. Springer.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010.

Sylvia Springorum, Sabine Schulte im Walde, and Jason Utt. 2013. Detecting polysemy in hard and soft cluster analyses of German preposition vector spaces. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 632–640.

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 71–78.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Automatic classification of English verbs using rich syntactic features. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 769–774.

Lin Sun, Diana McCarthy, and Anna Korhonen. 2013. Diathesis alternation approximation for verb clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Short Papers*, pages 736–741.

Robert Swier and Suzanne Stevenson. 2005. Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 883–890.

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).

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Ivan Titov and Alexandre Klementiev. 2012. A Bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 12–22.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 74–82.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the Workshop on Human Language Technology*, pages 266–271.