Supervised Learning of a Probabilistic Lexicon of Verb Semantic Classes

Yusuke Miyao
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
yusuke@is.s.u-tokyo.ac.jp

Jun'ichi Tsujii
University of Tokyo
University of Manchester
National Center for Text Mining
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
tsujii@is.s.u-tokyo.ac.jp

Abstract

The work presented in this paper explores a supervised method for learning a probabilistic model of a lexicon of VerbNet classes. We intend for the probabilistic model to provide a probability distribution of verb-class associations, over known and unknown verbs, including polysemous words. In our approach, training instances are obtained from an existing lexicon and/or from an annotated corpus, while the features, which represent syntactic frames, semantic similarity, and selectional preferences, are extracted from unannotated corpora. Our model is evaluated in type-level verb classification tasks: we measure the prediction accuracy of VerbNet classes for unknown verbs, and also measure the dissimilarity between the learned and observed probability distributions. We empirically compare several settings for model learning, while we vary the use of features, source corpora for feature extraction, and disambiguated corpora. In the task of verb classification into all VerbNet classes, our best model achieved a 10.69% error reduction in the classification accuracy, over the previously proposed model.

1 Introduction

Lexicons are invaluable resources for semantic processing. In many cases, lexicons are necessary to restrict a set of semantic classes to be assigned to a word. In fact, a considerable number of works on semantic processing implicitly or explicitly presupposes the availability of a lexicon, such as in word sense disambiguation (WSD) (McCarthy et al., 2004), and in token-level verb class disambiguation (Lapata and Brew, 2004; Girju et al., 2005; Li and Brew, 2007; Abend et al., 2008). In other words, those methods are heavily dependent on the availability of a semantic lexicon. Therefore, recent research efforts have invested in developing semantic resources, such as WordNet (Fellbaum, 1998), FrameNet (Baker et al., 1998), and VerbNet (Kipper et al., 2000; Kipper-Schuler, 2005), which greatly advanced research in semantic processing. However, the construction of such resources is expensive, and it is unrealistic to presuppose the availability of full-coverage lexicons; this is the case because unknown words always appear in real texts, and word-semantics associations may vary (Abend et al., 2008).

This paper explores a method for the supervised learning of a probabilistic model for the VerbNet lexicon. We target the automatic classification of arbitrary verbs, including polysemous verbs, into all VerbNet classes; further, we target the estimation of a probabilistic model, which represents the saliences of verb-class associations for polysemous verbs. In our approach, an existing lexicon and/or an annotated corpus are used as the training data. Since VerbNet classes are designed to represent the distinctions in the syntactic frames that verbs can take, features, representing the statistics of syntactic frames, are extracted from the unannotated corpora. Additionally, as the classes represent semantic commonalities, semantically inspired features, like distributionally similar words, are used. These features can be considered as a generalized representation of verbs, and we expect that the obtained probabilistic model predicts VerbNet classes of the unknown words.

Our model is evaluated in two tasks of type-level verb classification: one is the classification of monosemous verbs into a small subset of the classes, which was studied in some previous works (Joanis and Stevenson, 2003; Joanis et al., 2008). The other task is the classification of all verbs into the full set of VerbNet classes, which has not yet

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1328–1337,
Singapore, 6-7 August 2009.

been attempted. In the experiments, training instances are obtained from VerbNet and/or SemLink (Loper et al., 2007), while features are extracted from the British National Corpus or from Wall Street Journal. We empirically compare several settings for model learning by varying the set of features, the source domain and the size of a corpus for feature extraction, and the use of the token-level statistics obtained from a manually disambiguated corpus. We also provide the analysis of the remaining errors, which will lead us to further improve the supervised learning of a probabilistic semantic lexicon.

Supervised methods for automatic verb classification have been extensively investigated (Stevenson et al., 1999; Stevenson and Merlo, 1999; Merlo and Stevenson, 2001; Stevenson and Joanis, 2003; Joanis and Stevenson, 2003; Joanis et al., 2008). However, their focus has been limited to a small subset of verb classes, and a limited number of monosemous verbs. The main contributions of the present work are: i) to provide empirical results for the automatic classification of all verbs, including polysemous ones, into all VerbNet classes, and ii) to empirically explore the effective settings for the supervised learning of a probabilistic lexicon of verb semantic classes.

2 Background

2.1 Verb lexicon

Levin's (1993) work on verb classification has broadened the field of computational research that concerns the relationships between the syntactic and semantic structures of verbs. The principal idea behind the work is that the meanings of verbs can be identified by observing possible syntactic frames that the verbs can take. In other words, with the knowledge of syntactic frames, verbs can be semantically classified. This idea provided the computational linguistics community with criteria for the definition and the classification of verb semantics; it has subsequently resulted in the research of the induction of verb classes (Korhonen and Briscoe, 2004), and the construction of a verb lexicon based on Levin's criteria.

VerbNet (Kipper et al., 2000; Kipper-Schuler, 2005) is a lexicon of verbs organized into classes that share the same syntactic behaviors and semantics. The design of classes originates from Levin (1993), though the design has been considerably reorganized and extends beyond the original classification. The classes therefore cover more English verbs, and the classification should be more consistent (Korhonen and Briscoe, 2004; Kipper et al., 2006).

43 Emission
- **43.1 Light Emission**
  beam, glow, sparkle, ...
- **43.2 Sound Emission**
  blare, chime, jangle, ...

...

44 *Destroy*
annihilate, destroy, ravage, ...

45 Change of State

...

47 Existence
- **47.1 *Exist***
  exist, persist, remain, ...
- **47.2 Entity-Specific Modes Being**
  bloom, breathe, foam, ...
- **47.3 Modes of Being with Motion**
  jiggle, sway, waft, ...

...

Figure 1: VerbNet classes

43.2 Sound Emission
```
Theme V
Theme V P:loc Location
P:loc Location V Theme
there V Theme P:loc Location
Agent V Theme
Theme V Oblique
Location V with Theme
```

47.3 Modes of Being with Motion
```
Theme V
Theme V P:loc Location
P:loc Location V Theme
there V Theme
Agent V Theme
```

Figure 2: Syntactic frames for VerbNet classes

The current version of VerbNet includes 270 classes.[1] Figure 1 shows a part of the classes of VerbNet. The top-level categories, e.g. **Emission** and ***Destroy***, represent a coarse classification of verb semantics. They are further classified into verb classes, each of which expresses a group of verbs sharing syntactic frames. Figure 2 shows an excerpt from VerbNet, which represents the possible syntactic frames for the **Sound Emission** class, including "*chime*" and "*jangle*," and the **Modes of Being with Motion** class, including "*jiggle*" and "*waft*." In this figure, each line represents a syntactic frame, where `Agent`,

[1] Throughout this paper, we refer to VerbNet 2.3. Subclasses are ignored in this work, following the setting of Abend et al. (2008).

...the walls still shook;VN=47.3 and an evacuation alarm blared;VN=43.2 outside.

Suddenly the woman begins;VN=55.1 swaying ;VN=47.3 and then ...

Figure 3: An excerpt from SemLink

Theme, and Location indicate the thematic roles, V denotes a verb, and P specifies a preposition. P:loc defines locative prepositions such as: "*in*" and "*at*." For example, the second syntactic frame of **Sound Emission**, i.e., Theme V P:loc Location, corresponds to the following sentence:

1. The coins *jangled* in my pocket.

Theme corresponds to "*the coins*," V to "*jangled*," P:loc to "*in*," and Location to "*my pocket*."

While VerbNet provides associations between verbs and semantic classes, SemLink (Loper et al., 2007) additionally provides mappings among VerbNet, FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), and WordNet (Fellbaum, 1998). Since FrameNet and PropBank include annotated instances of sentences, SemLink can be considered as a corpus annotated with VerbNet classes. Figure 3 presents some annotated sentences obtained from SemLink. For example, the annotation "blared;VN=43.2" indicates that the occurrence of "blare" in this context is classified as **Sound Emission**.

2.2 Related work

There has been much research effort invested in the automatic classification of verbs into lexical semantic classes, in a supervised or unsupervised way. The present work inherits the spirit of the supervised approaches to verb classification (Stevenson et al., 1999; Stevenson and Merlo, 1999; Merlo and Stevenson, 2001; Stevenson and Joanis, 2003; Joanis and Stevenson, 2003; Joanis et al., 2008). Our learning framework basically follows the above listed works: features are obtained from an unannotated (automatically parsed) corpus, and gold verb-class associations are used as training instances for machine learning classifiers, such as decision trees and support vector machines. However, those works targeted a small subset of Levin classes, and a limited number of monosemous verbs; for example, Merlo and Stevenson (2001) studied three classes and 59 verbs, and Joanis et al. (2008) focused on 14 classes and 835 verbs. Although these works provided a theoretical framework for supervised verb classification, their results were not readily available for practical applications, because of the limitation in the coverage of the targeted classes/verbs on real texts. On the contrary, we target the classification of arbitrary verbs, including polysemous verbs, into all VerbNet classes (270 in total). In this realistic situation, we will empirically compare settings for model learning, in order to explore effective conditions to obtain better models.

Another difference from the aforementioned works is that we aim at obtaining a probabilistic model, which represents *saliences* of classes of polysemous verbs. Lapata and Brew (2004) and Li and Brew (2007) focused on this issue, and described methods for inducing probabilities of verb-class associations. The obtained probabilistic model was intended to be incorporated into a token-level disambiguation model. Their methods claimed to be unsupervised, meaning that the induction of a probabilistic lexicon did not require any hand-annotated corpora. In fact, however, their methods relied on the existence of a full-coverage lexicon, both in training and running time. In their methods, a lexicon was necessary for restricting possible classes to which each word belongs. Since most verbs are associated with only a couple of classes, such a restriction significantly reduces the search space, and the problem becomes much easier to solve. This presupposition is implicitly or explicitly used in other semantic disambiguation tasks (McCarthy et al., 2004), but it is unrealistic for practical applications.

Clustering methods have also been extensively researched for verb classification (Stevenson and Merlo, 1999; Schulte im Walde, 2000; McCarthy, 2001; Korhonen, 2002; Korhonen et al., 2003; Schulte im Walde, 2003). The extensive research is in large part due to the intuition that the set of classes could not be fixed beforehand. In particular, it is often problematic to define a static set of semantic classes. However, it is reasonable to assume that the set of VerbNet classes is fixed, because Levin-type classes are more static than ontological classes, like in WordNet synsets. Therefore, we can apply supervised classification methods to our task. It is true that the current VerbNet classes are imperfect and require revisions, but in this work we adopt them as they are, because as

time advances, more stable classifications will become available.

The problem focused in this work has a close relationship with automatic thesaurus/ontology expansion. In fact, we evaluate our method in the task of automatic verb classification, which can be considered as lexicon expansion. The most prominent difference of the present work from thesaurus/ontology expansion is that the number of classes is much smaller in our problem, and the set of verb classes can be assumed to be fixed. These characteristics indicate that our problem is easier and more well-defined than is the case for automatic thesaurus/ontology expansion.

Supervised approaches to token-level verb class disambiguation have recently been addressed (Girju et al., 2005; Abend et al., 2008), largely owing to SemLink. Their approaches fundamentally follow traditional supervised WSD methods: extracting features representing the context in which the target word appears, and training a classification model with an annotated corpus. While those works achieved an impressive accuracy (more than 95%), the results may not necessarily indicate the method's effectiveness; rather, it may imply the importance of a lexicon. In fact, these works restrict their target to verb tokens, in which the correct class exists in a given lexicon, and they only consider candidate classes that are registered in the lexicon. This setting reduces the ambiguity significantly, and the problem becomes much easier to handle; for example, approximately half of verb tokens are monosemous in their setting. Thus, a simple baseline achieves very high accuracy figures. However, in our preliminary experiment on token-level verb classification with unknown verbs, we found that the accuracy for unknown verbs (i.e., lemmas not included in the VerbNet lexicon) is catastrophically low. This indicates that VerbNet and SemLink are insufficient for unknown verbs, and that we cannot expect the availability of a full-coverage lexicon in the real world. Instead of a static lexicon, our probabilistic model is intended to be used as a prior distribution for the token-level disambiguation, as in Lapata and Brew (2004)'s model.

3 A probabilistic model for verb semantic classes

In this work, supervised learning is applied to the probabilistic modeling of a lexicon of verb semantic classes. We do not presuppose the existence of a full-coverage lexicon; instead, we use an existing lexicon for the training data. Combined with features extracted from unannotated corpora, a probabilistic model is learned from the existing lexicon. Like other supervised learning applications, our probabilistic lexicon can predict classes for words that are not included in the original lexicon.

Our model is defined in the following way. We assume that the set, C, of verb classes is fixed, while a set of verbs is unfixed. With this assumption, probabilistic modeling can be reduced to a classification problem. Specifically, the goal is to obtain a probability distribution, $p(c|v)$, of verb class $c \in C$ for a given verb (lemma) v. We can therefore apply well-known supervised learning methods to estimate $p(c|v)$.

This probability is modeled in the form of a log-linear model.

$$p(c|v) = \frac{1}{Z} \exp\left(\sum_i \lambda_i f_i(c, v)\right),$$

where $f_i(c, v)$ are features that represent characteristics of c and v, and λ_i are model parameters that express weights of the corresponding features.

Model parameters can be estimated when *training instances*, i.e., pairs $\langle c, v\rangle$, and *features*, $f_i(c, v)$, for each instance are given. Therefore, what we have to do is to prepare the training instances $\langle c, v\rangle$, and effective features $f_i(c, v)$ that contribute to the better estimation of probabilities. In token tagging tasks, both training instances and features are extracted from annotated corpora. However, since our goal is the probabilistic modeling of a lexicon, we have to determine how to derive the training instances and features for lexicon entries, to be discussed in the next section.

For the parameter estimation of log-linear models, we applied the stochastic gradient descent method. A hyperparameter for l_2-regularization was tuned to minimize the KL-divergence (see Section 4.4) for the development set.

4 Experiment design

In this work, we empirically compare several settings for the learning of the above probabilistic model, in the two tasks of automatic verb classification. In what follows, we explain the training/test data, corpora for extracting features, and the design of the features and evaluation tasks. The measures for evaluation are also introduced.

1	sound_emission-43.2	chime
0.5	sound_emission-43.2	blare
0.5	manner_speaking-37.3	blare
0.5	modes_of_being_with_motion-47.3	sway
0.5	urge-58.1	sway

1	sound_emission-43.2	chime
0.7	sound_emission-43.2	blare
0.3	manner_speaking-37.3	blare
0.6	modes_of_being_with_motion-47.3	sway
0.4	urge-58.1	sway

Figure 4: Training instances obtained from VerbNet (upper) and VerbNet+SemLink (lower)

4.1 Data

As our goal is the supervised learning of a lexicon of verb semantic classes, VerbNet is used as the training/test data. In addition, since we aim at representing the saliences of verb-class associations with probabilities, the gold probabilities are necessary. For this purpose, we count the occurrences of each verb-class association in the VerbNet-PropBank token mappings in the subset of the SemLink corresponding to sections 2 through 21 of Penn Treebank (Marcus et al., 1994). Frequency counts are normalized for each lemma, with the Laplace smoothing (the parameter is 0.5).

In this work, we compare the two settings for creating training instances. By comparing the results of these settings, we evaluate the necessity of an annotated corpus for learning a probabilistic lexicon of verb semantic classes.

VerbNet We collect all $\langle c, v \rangle$ pairs registered in VerbNet. For each v, all of the associated classes are assigned equal weights (see the upper part of Figure 4).

VerbNet+SemLink Each pair $\langle c, v \rangle$ in VerbNet is weighted by the normalized frequency obtained from SemLink (see the lower part of Figure 4).

Because VerbNet classes represent groups of syntactic frames, and it is impossible to guess the verb class by referring to only one occurrence in a text, it is necessary to have statistics over a sufficient amount of a corpus. Hence, features are extracted from a large unannotated corpus. In this paper, we use the following two corpora:

WSJ Wall Street Journal newspaper articles (around 40 million words).

BNC British National Corpus, which is a balanced corpus of around 100 million words.

In addition to the variance of the corpus domains, we vary the size of the corpus to observe the effect of increasing the corpus size. These corpora are automatically parsed by Enju 2.3.1 (Miyao and Tsujii, 2008), and the features are extracted from the parsing results.

4.2 Features

Levin-like classes, including VerbNet, are designed to represent distinctions in syntactic frames and alternations. Hence, if we were given the perfect knowledge of the possible syntactic frames, verbs can be classified into the correct classes almost perfectly (Dorr and Jones, 1996). Previous works thus proposed features that express the corpus statistics of syntactic frames. However, class boundaries are subtle in some cases; several classes share syntactic frames with each other to a large extent.

For example, the classes shown in Figure 2 have very similar syntactic frames. The difference is indicated in the last two frames of **Sound Emission**, although they appear much less frequently in real texts. Therefore, it is difficult to accurately capture the distinctions between these classes, if we are only provided with the statistics of the syntactic frames that appear in real texts. In this case, however, it is easy to observe that the verbs of these classes have different selectional preferences; that is, the `Theme` of **Sound Emission** verbs would be objects that make sounds, while the `Theme` of **Modes of Being with Motion** is likely to be objects that move.[2] Although Levin's classification initially focused on syntactic alternations, the resulting classes represent some semantic commonalities. Hence, it would be reasonable to design features that capture such semantic characteristics.

In this work, we re-implemented the following features proposed by Joanis et al. (2008) as the starting point.

Syntactic slot Features to count the occurrences of each syntactic slot, such as subject, object, and prepositional phrases. For the subject slot, we also count its transitive and intransitive usages separately. Additionally, we count the appearances of reflexive pronouns and semantically empty constituents (*it* and

[2] Syntactic frames in VerbNet include specifications of selectional preferences, such as *animate* and *place*, although we do not explicitly use them, because it is not apparent to determine the members of these semantic classes.

Syntactic slot	`subj:0.885` `intrans-subj:0.578`
Slot overlap	`overlap-subj-obj:0.299` `overlap-obj-in:0.074`
Tense, voice, aspect	`pos-VBG:0.307` `pos-VBD:0.290`
Animacy	`anim-subj:0.244` `anim-obj:0.057`
Slot POS	`subj-PRP:0.270` `subj-NN:0.270`
Syntactic frame	`NP_V:0.326` `NP_V_NP:0.307`
Similar word	`sim-rock:0.090` `sim-swing:0.083`
Slot class	`subj-C82:0.219` `obj-C12:0.081`

Figure 5: Example of features for "sway"

there). Differently from Joanis et al. (2008), we consider non-nominal arguments, such as sentential and adjectival complements.

Slot overlap Features to measure the overlap in words (lemmas) between two syntactic slots of the verb. They are intended to approximate argument alternations, such as the ergative alternation. For example, for the alternation "*The sky cleared*"/"*The clouds cleared from the sky*," a feature to indicate the overlap between the subject slot and the *from* slot is added (Joanis et al., 2008). The value of this feature is computed by the method of Merlo and Stevenson (2001).

Tense, voice, aspect Features to approximate the tendency of the tense, voice, and aspect of the target verb. The Penn Treebank POS tags for verbs (VB, VBP, VBZ, VBG, VBD, and VBN) are counted. In addition, included are the frequency of the co-occurrences with an adverb or an auxiliary verb, and the count of usages as a noun or an adjective.

Animacy Features to measure the frequency of animate arguments for each syntactic slot. Personal pronouns except *it* are counted as animate, following Joanis et al. (2008), while named entity recognition was not used.

Examples of these features are shown in Figure 5. For details, refer to Joanis et al. (2008).

The above features mainly represent syntactic behaviors of target verbs. Since our target classes are broader than in the previous works, we further enhance the syntactic features. Additionally, as discussed above, semantically motivated features may present strong clues to distinguish among syntactically similar classes. We therefore include the following four types of feature; the first two are syntactic, while the other two are intended to capture semantic characteristics:

Slot POS In addition to the syntactic slot features, we add features that represent a combination of a syntactic slot and the POS of its head word. Since VerbNet includes extended classes that take verbal and adjectival arguments, the POSs of arguments would provide a strong clue to discriminate among these syntactic frames.

Syntactic frame The number of arguments and their syntactic categories. This feature was mentioned as a baseline in Joanis et al. (2008), but we include it in our model.

Similar word Similar words (lemmas) to the target verb. Similar words are automatically obtained from a corpus (the same corpus as used for feature extraction) by Lin (1998)'s method. This feature is motivated by the hypothesis that distributionally similar words tend to be classified into the same class. Because Lin's method is based on the similarity of words in syntactic slots, the obtained similar words are expected to represent a verb class that share selectional preferences.

Slot class Semantic classes of the head words of the arguments. This feature is also intended to approximate selectional preferences. The semantic classes are obtained by clustering nouns, verbs, and adjectives into 200, 100, and 50 classes respectively, by using the k-medoid method with Lin (1998)'s similarity.

Figure 5 shows an example of the features for "sway," extracted from the BNC corpus.[3] Feature values are defined as relative frequencies for each lemma; while, for similar word features, feature values are weighted by Lin's similarity measure.

4.3 Tasks

We evaluate our model in the tasks of automatic verb classification (a.k.a. lexicon expansion): given gold verb-class associations for some set of verbs, we predict the classes for unknown

[3] "C82" and "C12" are automatically assigned cluster names.

Verb class	Levin class number
Recipient	13.1, 13.3
Admire	31.2
Amuse	31.1
Run	51.3.2
Sound Emission	43.2
Light and Substance Emission	43.1, 43.4
Cheat	10.6
Steal and *Remove*	10.5, 10.1
Wipe	10.4.1, 10.4.2
Spray/Load	9.7
Fill	9.8
Other Verbs of Putting	9.1–6
Change of State	45.1–4
Object Drop	26.1, 26.3, 26.7

Table 1: 14 classes used in Joanis et al. (2008) and their corresponding Levin class numbers

verbs. While our main target is the full set of VerbNet classes, we also show results for the task studied in the previous work.

14-class task The task to classify (almost) monosemous verbs into 14 classes. Refer to Table 1 for the definition of the 14 classes. Following Joanis et al. (2008)'s task definition, we removed verbs that belong to multiple classes in these 14 classes, and also removed *overly polysemous* verbs (in our experiment, verb-class associations that have the relative frequency that is less than 0.5 in SemLink are removed). For each class, member verbs are randomly split into 50% (training), 25% (development), and 25% (final test) sets.

All-class task The task to classify all target verbs into 268 classes.[4] Any verbs that did not occur at least 100 times in the BNC corpus were removed.[5] The remaining verbs (2517 words) are randomly split into 80% (training), 10% (development), and 10% (final test) sets, under the constraint that at least one instance for each class is included in the training set.[6]

4.4 Evaluation measures

For the 14-class task, we simply measure the classification accuracy. However, the evaluation in the all-class task is not trivial, because verbs may be assigned multiple classes.

Since our purpose is to obtain a probabilistic model rather than to classify monosemous verbs, the evaluation criterion should be sensitive to the probabilistic distribution on the test data. In this paper, we adopt two evaluation measures. One is the *top-N weighted accuracy*; we count the number of correct pairs $\langle c, v \rangle$ in the N-best outputs from the model (where N is the number of gold classes for each lemma), where each count is weighted by the relative frequency (i.e., the counts in SemLink) of the pair in the test set. For example, in the case for "blare" in Figure 4, if the model states that **Sound Emission** has the largest probability, we get 0.7 points. If **Manner Speaking** has the largest probability, we instead obtain 0.3 points. Intuitively, the score is higher when the model presents larger probabilities to classes with higher relative frequencies. This measure is similar to the top-N precision in information retrieval; it evaluates the ranked output by the model. It is intuitively interpretable, but is insufficient for evaluating the quality of probability distributions.

The other measure is *KL-divergence*, which is popularly used for measuring the dissimilarity between two probability distributions. This is defined as follows:

$$KL(p||q) = \sum_x p(x)\log(p(x)) - p(x)\log(q(x)).$$

In the experiments, this measure is applied, with the assumption that p is the relative frequency of $\langle c, v \rangle$ in the test set, and that q is the estimated probability distribution. Although the KL-divergence is not a true distance metric, it is sufficient for measuring the fitting of the estimated model to the true distribution. We report the KL-divergence averaged over all verbs in the test set. Since this measure indicates a dissimilarity, a smaller value is better. When p and q are equivalent, $KL(p||q) = 0$.

5 Experimental results

Table 2 shows the accuracy obtained for the 14-class task. The first column denotes the incorporated features ("Joanis et al.'s features" or "All features"), and the sources of the features ("WSJ" or "BNC"). The two baseline results are also given: "Baseline (*random*)" indicates that classes are randomly output, and "Baseline (*majority*)" indicates

[4] Two classes (**Being Dressed** and **Debone**) are not used in the experiments because no lemmas belonged to these classes after filtering by the frequency in BNC.

[5] This is the same preprocessing as Joanis et al. (2008), although we use VerbNet, while Joanis et al. (2008) used the original Levin classifications.

[6] Because polysemous verbs belong to multiple classes, the class-wise data split was not adopted for the all-class task.

	Accuracy
Baseline (*random*)	7.14
Baseline (*majority*)	26.47
Joanis et al.'s features/WSJ	56.86
Joanis et al.'s features/BNC	64.22
All features/WSJ	60.29
All features/BNC	68.14

Table 2: Accuracy for the 14-class task

	Accuracy	KL
Baseline (*random*)	0.37	—
Baseline (*majority*)	8.69	—
Joanis et al.'s features/WSJ	30.26	3.65
Joanis et al.'s features/BNC	35.66	3.32
All features/WSJ	34.07	3.37
All features/BNC	42.54	2.99

Table 3: Accuracy and KL-divergence for the all-class task (the VerbNet+SemLink setting)

that the majority class (i.e., the class that has the largest number of member verbs) is output to every lemma. While these figures cannot be compared directly to the previous works due to the difference in the preprocessing, Joanis et al. (2008) achieved 58.4% accuracy for the 14-class task. Table 3 and 4 present the results for the all-class task. Table 3 gives the accuracy and KL-divergence achieved by the model trained with the VerbNet+SemLink training instances, while Table 4 presents the same measures by the training instances created from VerbNet only.

Our models performed substantially better on both tasks than the baseline models. The results also proved that the features we proposed in this paper contributed to the further improvement of the model from Joanis et al. (2008). In the all-class task with the VerbNet+SemLink setting, our features achieved 10.69% error reduction in the accuracy over Joanis et al. (2008)'s features. Another interesting fact is that the model with BNC consistently outperformed the model with WSJ. This outcome is somewhat surprising, provided that the relative frequencies in the training/test sets are created from the WSJ portion of SemLink. The reason for this is independent of the corpus size, as will be shown below. When comparing Table 3 and 4, we can see that using SemLink statistics resulted in a slightly better model. This result is predictable, because the evaluation measures are sensitive to the relative frequencies estimated from SemLink. However, the difference remained small. In both of the tasks and the evaluation measures, the best model was achieved when we use

	Accuracy	KL
Baseline (*random*)	0.37	—
Baseline (*majority*)	8.69	—
Joanis et al.'s features/WSJ	29.65	3.67
Joanis et al.'s features/BNC	35.78	3.34
All features/WSJ	34.53	3.40
All features/BNC	42.38	3.02

Table 4: Accuracy and KL-divergence for the all-class task (the VerbNet only setting)

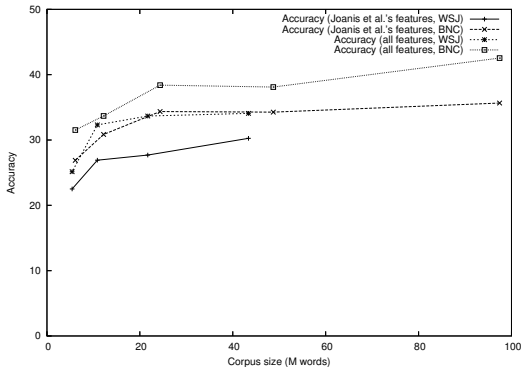

Figure 6: Corpus size vs. accuracy

all the features extracted from BNC, and create training instances from VerbNet+SemLink.

Figure 6 and 7 plot the accuracy and KL-divergence against the size of the unannotated corpus used for feature extraction. The result clearly indicates that the learning curve still grows at the corpus size with 100 million words (especially for the all features + BNC setting), which indicates that better models are obtained by increasing the size of the unannotated corpora.

Therefore, we can claim that the differences between the domains and the size of the unannotated corpora are more influential than the availability of the annotated corpora. This indicates that learning only from a lexicon would be a viable solution, when a token-disambiguated corpus like SemLink is unavailable.

Table 5 shows the contribution of each feature group. BNC is used for feature extraction, and VerbNet+SemLink is used for the creation of training instances. The results demonstrated the effectiveness of the slot POS features, and in particular, for the all-class task, most likely because VerbNet covers verbs that take non-nominal arguments. Additionally, the similar word features contributed equally or more in both of the tasks. This result suggests that we were reasonable in hypothesizing that distributionally similar words tend to be clas-

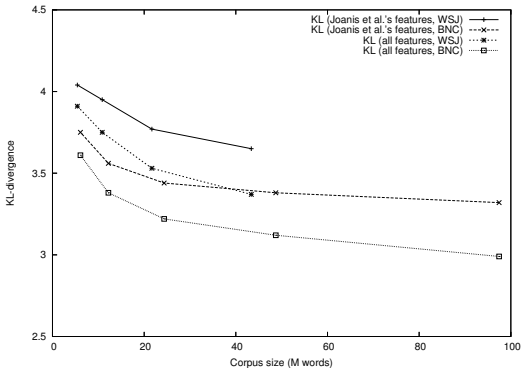

Figure 7: Corpus size vs. KL-divergence

	14-classes Accuracy	All classes Accuracy	KL
Baseline (*random*)	7.14	0.37	—
Baseline (*majority*)	26.47	8.69	—
Joanis et al.'s features	64.22	35.66	3.32
+ Slot POS	66.67	38.77	3.18
+ Syntactic frame	64.71	35.99	3.29
+ Similar word	68.14	37.88	3.10
+ Slot class	64.71	36.51	3.26
All features	68.14	42.54	2.99

Table 5: Contribution of features

sified into the same class. Slot classes also contributed to a slight improvement, indicating that selectional preferences are effective clues for predicting VerbNet classes. The result of the "All features" model for the all-class task attests that these features worked collaboratively, and using them all resulted in a considerably better model.

From the analysis of the confusion matrix for the outputs by our best model, we identified several reasons for the remaining misclassification errors. A major portion of the errors were caused by confusing the classes that take the same prepositions. Examples of these errors include:

- **Other Change of State** verbs were misclassified into the ***Butter*** class: "embalm," "laminate." (they take "with" phrases)
- **Judgement** verbs were misclassified into the **Characterize** class: "acclaim," "hail." (they take "as" phrases)

Since prepositions are strong features for automatic verb classification (Joanis et al., 2008), the classes that take the same prepositions remained confusing. The discovery of the features to discriminate among these classes would be crucial for further improvement.

Another major error is in classifying verbs into **Other Change of State**. Examples include:

- ***Amuse*** verbs: "impair," "recharge."
- ***Herd*** verbs: "aggregate," "mass."

Because **Other Change of State** is one of the biggest classes, supervised learning tends to place a high probability to this class. Therefore, when strong clues do not exist, verbs tend to be misclassified into this class. In addition, this class is not syntactically/semantically homogeneous, and is likely to introduce noise in the machine learning classifier. A possible solution to this problem would be to exclude this class from the classification, and to process the class separately.

6 Conclusions

We presented a method for the supervised learning of a probabilistic model for a lexicon of VerbNet classes. By combining verb-class associations from VerbNet and SemLink, and features extracted from a large unannotated corpus, we could successfully train a log-linear model in a supervised way. The experimental results attested to our success that features proposed in this paper worked effectively in obtaining a better probability distribution. Not only syntactic features, but also semantic features were shown to be effective. While each of these features could increase the accuracy, they collaboratively contributed to a large improvement. In the all-class task, we obtained 10.69% error reduction in the classification accuracy over Joanis et al. (2008)'s model. We also observed the trend that a larger corpus for feature extraction led to a better model, indicating that a better model will be obtained by increasing the size of an unannotated corpus.

We could identify the effective features and settings for this problem, but the classification into all VerbNet classes remained challenging. One possible direction for this research topic would be to use our model for the semi-automatic construction of verb lexicons, with the help of human curation. However, there is also a demand for exploring other types of features that can discriminate among confusing classes.

Acknowledgments

This work was partially supported by Grant-in-Aid for Specially Promoted Research and Grant-in-Aid for Young Scientists (MEXT, Japan).

References

Omri Abend, Roi Reichart, and Ari Rappoport. 2008. A supervised algorithm for verb disambiguation into VerbNet classes. In *Proceedings of COLING 2008*, pages 9–16.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL 1998*.

Bonnie J. Dorr and Doug Jones. 1996. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of COLING-96*, pages 322–327.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Roxana Girju, Dan Roth, and Mark Sammons. 2005. Token-level disambiguation of VerbNet classes. In *The Interdisciplinary Workshop on Verb Features and Verb Classes*.

Eric Joanis and Suzanne Stevenson. 2003. A general feature space for automatic verb classification. In *Proceedings of EACL 2003*, pages 163–170.

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

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of 17th National Conference on Artificial Intelligence*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC 2006*.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Department, University of Pennsylvania.

Anna Korhonen and Ted Briscoe. 2004. Extended lexical-semantic classification of English verbs. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*.

Anna Korhonen, Yuval Krymolowski, and Zvika Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of ACL 2003*.

Anna Korhonen. 2002. Semantically motivated subcategorization acquisition. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, pages 51–58.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–75.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Juanguo Li and Chris Brew. 2007. Disambiguating Levin verbs using untagged data. In *Proceedings of RANLP 2007*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 1998*.

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, Tilburg, the Netherlands*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of ACL 2004*.

Diana McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb-classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3):373–408.

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behavior. In *Proceedings of COLING 2000*, pages 747–753.

Sabine Schulte im Walde. 2003. Experiments on the choice of features for learning verb classes. In *Proceedings of EACL 2003*, pages 315–322.

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of CoNLL 2003*, pages 71–78.

Suzanne Stevenson and Paola Merlo. 1999. Automatic verb classification using grammatical features. In *Proceedings of EACL 1999*, pages 45–52.

Suzanne Stevenson, Paola Merlo, Natalia Kariaeva, and Kamin Whitehouse. 1999. Supervised learning of lexical semantic verb classes using frequency distributions. In *Proceedings of SigLex99: Standardizing Lexical Resources*, pages 15–22.