Proceedings of the Seventh Conference on Computational Natural Language Learning (CoNLL-03), pp. 71-78.

Semi-supervised Verb Class Discovery Using Noisy Features

Suzanne Stevenson and **Eric Joanis**
Department of Computer Science
University of Toronto
{suzanne,joanis}@cs.toronto.edu

Abstract

We cluster verbs into lexical semantic classes, using a general set of noisy features that capture syntactic and semantic properties of the verbs. The feature set was previously shown to work well in a supervised learning setting, using known English verb classes. In moving to a scenario of verb class discovery, using clustering, we face the problem of having a large number of irrelevant features for a particular clustering task. We investigate various approaches to feature selection, using both unsupervised and semi-supervised methods, comparing the results to subsets of features manually chosen according to linguistic properties. We find that the unsupervised method we tried cannot be consistently applied to our data. However, the semi-supervised approach (using a seed set of sample verbs) overall outperforms not only the full set of features, but the hand-selected features as well.

1 Introduction

Computational linguists face a lexical acquisition bottleneck, as vast amounts of knowledge about individual words are required for language technologies. Learning the argument structure properties of verbs—the semantic roles they assign and their mapping to syntactic positions—is both particularly important and difficult. A number of supervised learning approaches have extracted such information about verbs from corpora, including their argument roles (Gildea and Jurafsky, 2002), selectional preferences (Resnik, 1996), and lexical semantic classification (i.e., grouping verbs according to their argument structure properties) (Dorr and Jones, 1996; Lapata and Brew, 1999; Merlo and Stevenson, 2001; Joanis and Stevenson, 2003). Unsupervised or semi-supervised approaches have been successful as well, but have tended to be more restrictive, in relying on human filtering of the results (Riloff and Schmelzenbach, 1998), on the hand-selection of features (Stevenson and Merlo, 1999), or on the use of an extensive grammar (Schulte im Walde and Brew, 2002).

We focus here on extending the applicability of unsupervised methods, as in (Schulte im Walde and Brew, 2002; Stevenson and Merlo, 1999), to the lexical semantic classification of verbs. Such classes group together verbs that share both a common semantics (such as transfer of possession or change of state), and a set of syntactic frames for expressing the arguments of the verb (Levin, 1993; FrameNet, 2003). As such, they serve as a means for organizing complex knowledge about verbs in a computational lexicon (Kipper et al., 2000). However, creating a verb classification is highly resource intensive, in terms of both required time and linguistic expertise. Development of minimally supervised methods is of particular importance if we are to automatically classify verbs for languages other than English, where substantial amounts of labelled data are not available for training classifiers. It is also necessary to consider the probable lack of sophisticated grammars or text processing tools for extracting accurate features.

We have previously shown that a broad set of 220 noisy features performs well in supervised verb classification (Joanis and Stevenson, 2003). In contrast to Merlo and Stevenson (2001), we confirmed that a set of general features can be successfully used, without the need for manually determining the relevant features for distinguishing particular classes (cf. Dorr and Jones, 1996; Schulte im Walde and Brew, 2002). On the other hand, in contrast to Schulte im Walde and Brew (2002), we demonstrated that accurate subcategorization statistics are unnecessary (see also Sarkar and Tripasai, 2002).

By avoiding the dependence on precise feature extraction, our approach should be more portable to new languages. However, a general feature space means that most features will be irrelevant to any given verb discrimination task. In an unsupervised (clustering) scenario of verb class discovery, can we maintain the benefit of only needing noisy features, without the generality of the feature space leading to "the curse of dimensionality"? In supervised experiments, the learner uses class labels during the training stage to determine which features are relevant to the task at hand. In the unsupervised setting, the large number of potentially irrelevant features becomes a serious problem, since those features may mislead the learner.

Thus, the problem of dimensionality reduction is a key issue to be addressed in verb class discovery. In this paper, we report results on several feature selection approaches to the problem: manual selection (based on linguistic knowledge), unsupervised selection (based on an entropy measure among the features, Dash et al., 1997), and a semi-supervised approach (in which seed verbs are used to train a supervised learner, from which we extract the useful features). Although our motivation is verb class discovery, we perform our experiments on English, for which we have an accepted classification to serve as a gold standard (Levin, 1993). To preview our results, we find that, overall, the semi-supervised method not only outperforms the entire feature space, but also the manually selected subset of features. The unsupervised feature selection method, on the other hand, was not usable for our data.

In the remainder of the paper, we first briefly review our feature space and present our experimental classes and verbs. We then describe our clustering methodology, the measures we use to evaluate a clustering, and our experimental results. We conclude with a discussion of related work, our contributions, and future directions.

2 The Feature Space

Like others, we have assumed lexical semantic classes of verbs as defined in Levin (1993) (hereafter Levin), which have served as a gold standard in computational linguistics research (Dorr and Jones, 1996; Kipper et al., 2000; Merlo and Stevenson, 2001; Schulte im Walde and Brew, 2002). Levin's classes form a hierarchy of verb groupings with shared meaning and syntax. Our feature space was designed to reflect these classes by capturing properties of the semantic arguments of verbs and their mapping to syntactic positions. It is important to emphasize, however, that our features are extracted from part-of-speech (POS) tagged and chunked text only: there are no semantic tags of any kind. Thus, the features serve as approximations to the underlying distinctions among classes.

Here we briefly describe the features that comprise our feature space, and refer the interested reader to Joanis and Stevenson (2003) for details.

Features over Syntactic Slots (120 features)
One set of features encodes the frequency of the syntactic slots occurring with a verb (subject, direct and indirect object, and prepositional phrases (PPs) indexed by preposition), which collectively serve as rough approximations to the allowable syntactic frames for a verb. We also count fixed elements in certain slots (*it* and *there*, as in *It rains* or *There appeared a ship*), since these are part of the syntactic frame specifications for a verb.

In addition to approximating the syntactic frames themselves, we also want to capture regularities in the mapping of arguments to particular slots. For example, the location argument, *the truck*, is direct object in *I loaded the truck with hay*, and object of a preposition in *I loaded hay onto the truck*. These allowable alternations in the expressions of arguments vary according to the class of a verb. We measure this behaviour using features that encode the degree to which two slots contain the same entities—that is, we calculate the overlap in noun (lemma) usage between pairs of syntactic slots.

Tense, Voice, and Aspect Features (24 features)
Verb meaning, and therefore class membership, interacts in interesting ways with voice, tense, and aspect (Levin, 1993; Merlo and Stevenson, 2001). In addition to verb POS (which often indicates tense) and voice (passive/active), we also include counts of modals, auxiliaries, and adverbs, which are partial indicators of these factors.

The Animacy Features (76 features)
Semantic properties of the arguments that fill certain roles, such as animacy or motion, are more challenging to detect automatically. Currently, our only such feature is an extension of the animacy feature of Merlo and Stevenson (2001). We approximate the animacy of each of the 76 syntactic slots by counting both pronouns and proper noun phrases (NPs) labelled as "person" by our chunker (Abney, 1991).

3 Experimental Classes and Verbs

We use the same classes and example verbs as in the supervised experiments of Joanis and Stevenson (2003) to enable a comparison between the performance of the unsupervised and supervised methods. Here we describe the selection of the experimental classes and verbs, and the estimation of the feature values.

3.1 The Verb Classes

Pairs or triples of verb classes from Levin were selected to form the test pairs/triples for each of a number of separate classification tasks. These sets exhibit different contrasts between verb classes in terms of their semantic argument assignments, allowing us to evaluate our approach under a range of conditions. For example, some classes differ in both their semantic roles and frames, while others have the same roles in different frames, or different roles in the same frames.[1] Here we summarize the argument structure distinctions between the classes; Table 1 below lists the classes with their Levin class numbers.

Benefactive versus Recipient verbs.
Mary baked... a cake for Joan/Joan a cake.
Mary gave... a cake to Joan/Joan a cake.
These dative alternation verbs differ in the preposition and the semantic role of its object.

[1]For practical reasons, as well as for enabling us to draw more general conclusions from the results, the classes also could neither be too small nor contain mostly infrequent verbs.

<u>*Admire* versus *Amuse* verbs.</u>
I admire Jane. *Jane amuses me.*
These psychological state verbs differ in that the Experiencer argument is the subject of *Admire* verbs, and the object of *Amuse* verbs.

<u>*Run* versus Sound Emission verbs.</u>
*The kids ran in the room./*The room ran with kids.*
The birds sang in the trees./The trees sang with birds.
These activity verbs both have an Agent subject in the intransitive, but differ in the prepositional alternations they allow.

<u>*Cheat* versus *Steal* and *Remove* verbs.</u>
*I cheated... Jane of her money/*the money from Jane.*
*I stole... *Jane of her money/the money from Jane.*
These classes also assign the same semantic arguments, but differ in their prepositional alternants.

<u>*Wipe* versus *Steal* and *Remove* verbs.</u>
Wipe... the dust/the dust from the table/the table.
*Steal... the money/the money from the bank/*the bank.*
These classes generally allow the same syntactic frames, but differ in the possible semantic role assignment. (Location can be the direct object of *Wipe* verbs but not of *Steal* and *Remove* verbs, as shown.)

<u>*Spray/Load* versus *Fill* versus Other Verbs of Putting (several related Levin classes).</u>
I loaded... hay on the wagon/the wagon with hay.
*I filled... *hay on the wagon/the wagon with hay.*
*I put... hay on the wagon/*the wagon with hay.*
These three classes also assign the same semantic roles but differ in prepositional alternants. Note, however, that the options for *Spray/Load* verbs overlap with those of the other two types of verbs.

<u>Optionally Intransitive: *Run* versus Change of State versus "Object Drop".</u>
The horse raced./The jockey raced the horse.
The butter melted./The cook melted the butter.
The boy played./The boy played soccer.
These three classes are all optionally intransitive but assign different semantic roles to their arguments (Merlo and Stevenson, 2001). (Note that the Object Drop verbs are a superset of the Benefactives above.)

For many tasks, knowing exactly what PP arguments each verb takes may be sufficient to perform the classification (cf. Dorr and Jones, 1996). However, our features do not give us such perfect knowledge, since PP arguments and adjuncts cannot be distinguished with high accuracy. Using our simple extraction tools, for example, the PP_{for} argument in *I admired Jane for her honesty* is not distinguished from the PP_{for} adjunct in *I amused Jane for the money*. Furthermore, PP arguments differ in frequency, so that a highly distinguishing but rarely used alternant will likely not be useful. Indicators of PP usage are thus useful but not definitive.

Verb Class	Class Number	# Verbs
Benefactive	26.1, 26.3	35
Recipient	13.1, 13.3	27
Admire	31.2	35
Amuse	31.1	134
Run	51.3.2	79
Sound Emission	43.2	56
Cheat	10.6	29
Steal and *Remove*	10.5, 10.1	45
Wipe	10.4.1, 10.4.2	35
Spray/Load	9.7	36
Fill	9.8	63
Other V. of Putting	9.1–6	48
Change of State	45.1–4	169
Object Drop	26.1, 26.3, 26.7	50

Table 1: Verb classes (see Section 3.1), their Levin class numbers, and the number of experimental verbs in each (see Section 3.2).

3.2 Verb Selection

Our experimental verbs were selected as follows. We started with a list of all the verbs in the given classes from Levin, removing any verb that did not occur at least 100 times in our corpus (the BNC, described below). Because we make the simplifying assumption of a single correct classification for each verb, we also removed any verb: that was deemed excessively polysemous; that belonged to another class under consideration in our study; or for which the class did not correspond to the main sense.

Table 1 above shows the number of verbs in each class at the end of this process. Of these verbs, 20 from each class were randomly selected to use as training data for our supervised experiments in Joanis and Stevenson (2003). We began with this same set of 20 verbs per class for our current work. We then replaced 10 of the 260 verbs (4%) to enable us to have representative seed verbs for certain classes in our semi-supervised experiments (e.g., so that we could include *wipe* as a seed verb for the *Wipe* verbs, and *fill* for the *Fill* verbs). All experiments reported here were run on this same final set of 20 verbs per class (including a replication of our earlier supervised experiments).

3.3 Feature Extraction

All features were estimated from counts over the British National Corpus (BNC), a 100M word corpus of text samples of recent British English ranging over a wide spectrum of domains. Since it is a general corpus, we do not expect any strong overall domain bias in verb usage.

We used the chunker (partial parser) of Abney (1991) to preprocess the corpus, which (noisily) determines the NP subject and direct object of a verb, as well as the PPs potentially associated with it. Indirect objects are identified by a less sophisticated (and even noisier) method, simply assuming that two consecutive NPs after the verb constitute a double object frame. From these extracted slots, we calculate the features described in Section 2, yielding a vector of 220 normalized counts for each verb, which forms the input to our machine learning experiments.

4 Clustering and Evaluation Methods

4.1 Clustering Parameters

We used the hierarchical clustering command in Matlab, which implements bottom-up agglomerative clustering, for all our unsupervised experiments. In performing hierarchical clustering, both a vector distance measure and a cluster distance ("linkage") measure are specified. We used the simple Euclidean distance for the former, and Ward linkage for the latter. Ward linkage essentially minimizes the distances of all cluster points to the centroid, and thus is less sensitive to outliers than some other methods.

We chose hierarchical clustering because it may be possible to find coherent subclusters of verbs even when there are not exactly C good clusters, where C is the number of classes. To explore this, we can induce any number of clusters K by making a cut at a particular level in the clustering hierarchy. In the experiments here, however, we report only results for $K = C$, since we found no principled way of automatically determining a good cut-off. However, we did experiment with $K = 2C$ (as in Strehl et al., 2000), and found that performance was generally better (even on our R_{adj} measure, described below, that discounts oversplitting). This supports our intuition that the approach may enable us to find more consistent clusters at a finer grain, without too much fragmentation.

4.2 Evaluation Measures

We use three separate evaluation measures, that tap into very different properties of the clusterings.

4.2.1 Accuracy

We can assign each cluster the class label of the majority of its members. Then for all verbs v, consider v to be classified correctly if Class(v)=ClusterLabel(v), where Class(v) is the actual class of v and ClusterLabel(v) is the label assigned to the cluster in which v is placed. Then accuracy has the standard definition:[2]

$$Acc = \frac{\text{\#verbs correctly classified}}{\text{\#verbs total}}$$

Acc thus provides a measure of the usefulness in practice of a clustering—that is, if one were to use the clustering as a classification, this measure tells how accurate overall the class assignments would be. The theoretical maximum is, of course, 1. To calculate a random baseline, we evaluated 10,000 random clusterings with the same number of verbs and classes as in each of our experimental tasks. Because the Acc achieved depends on the precise size of clusters, we calculated mean Acc over the best scenario (with equal-sized clusters), yielding a conservative estimate (i.e., an upper bound) of the baseline. These figures are reported with our results in Table 2 below.

4.2.2 Adjusted Rand Measure

Accuracy can be relatively high for a clustering when a few clusters are very good, and others are not good. Our second measure, the adjusted Rand measure used by Schulte im Walde (2003), instead gives a measure of how consistent the given clustering is *overall* with respect to the gold standard classification. The formula is as follows (Hubert and Arabie, 1985):

$$R_{adj} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}/\binom{n}{2}}{\frac{1}{2}[\sum_i\binom{n_{i\cdot}}{2} + \sum_j\binom{n_{\cdot j}}{2}] - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}/\binom{n}{2}}$$

where n_{ij} is the entry in the contingency table between the classification and the clustering, counting the size of the intersection of class i and cluster j. Intuitively, R_{adj} measures the similarity of two partitions of data by considering agreements and disagreements between them—there is agreement, for example, if v_i and v_j from the same class are in the same cluster, and disagreement if they are not. It is scaled so that perfect agreement yields a value of 1, whereas random groupings (with the same number of groups in each) get a value around 0. It is therefore considered "corrected for chance," given a fixed number of clusters.[3]

In tests of the R_{adj} measure on some contrived clusterings, we found it quite conservative, and on our experimental clusterings it did not often attain values higher than .25. However, it is useful as a relative measure of goodness, in comparing clusterings arising from different feature sets.

4.2.3 Mean Silhouette

Acc gives an average of the individual goodness of the clusters, and R_{adj} a measure of the overall goodness, both with respect to the gold standard classes. Our final measure gives an indication of the overall goodness of the clusters purely in terms of their separation of the data, without

[2] Acc is equivalent to the weighted mean precision of the clusters, weighted according to cluster size.

As we have defined it, Acc necessarily generally increases as the number of clusters increases, with the extreme being at the number of clusters equal to the number of verbs. However, since we fix our number of clusters to the number of classes, the measure remains informative.

[3] In our experiments for estimating the Acc baseline, we indeed found a mean R_{adj} value of 0.00 for all random clusterings.

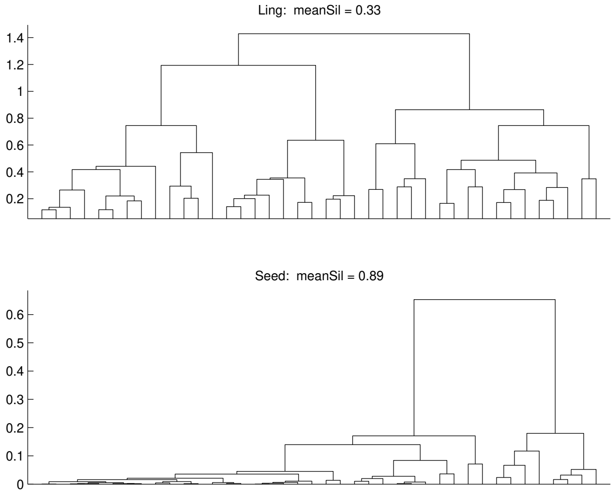

Figure 1: The dendrograms and $meanSil$ values for the 2-way *Wipe/Steal–Remove* task, using the Ling and Seed sets. The higher $meanSil$ (.89 vs. .33) reflects the better separation of the data.

regard to the target classes. We use $meanSil$, the mean of the silhouette measure from Matlab, which measures how distant a data point is from other clusters. Silhouette values vary from +1 to -1, with +1 indicating that the point is near the centroid of its own cluster, and -1 indicating that the point is very close to another cluster (and therefore likely in the wrong cluster). A value of 0 suggests that a point is not clearly in a particular cluster.

We calculate the mean silhouette of all points in a clustering to obtain an overall measure of how well the clusters are separated. Essentially, the measure numerically captures what we can intuitively grasp in the visual differences between the dendrograms of "better" and "worse" clusterings. (A dendrogram is a tree diagram whose leaves are the data points, and whose branch lengths indicate similarity of subclusters; roughly, shorter vertical lines indicate closer clusters.)

For example, Figure 1 shows two dendrograms using different feature sets (Ling and Seed, described in Section 5) for the same set of verbs from two classes. The Seed set has slightly lower values for Acc and R_{adj}, but a much higher value (.89) for $meanSil$, indicating a better separation of the data. This captures what is reflected in the dendrogram, in that very short lines connect verbs low in the tree, and longer lines connect the two main clusters.

The $meanSil$ measure is independent of the true classification, and could be high when the other dependent measures are low, or vice versa. However, it gives important information about the quality of a clustering: The other measures being equal, a clustering with a higher $meanSil$ value indicates tighter and more separated clusters, suggesting stronger inherent patterns in the data.

5 Experimental Results

We report here the results of a number of clustering experiments, using feature sets as follows: (1) the full feature space; (2) a manually selected subset of features; (3) unsupervised selection of features; and (4) semi-supervised selection, using a supervised learner applied to seed verbs to select the features.

For each type of feature set, we performed the same ten clustering tasks, shown in the first column of Table 2. These are the same tasks performed in the supervised setting of Joanis and Stevenson (2003). The 2- and 3-way tasks, and their motivation, were described in Section 3.1. Three multiway tasks explore performance over a larger number of classes: The 6-way task involves the *Cheat*, *Steal–Remove*, *Wipe*, *Spray/Load*, *Fill*, and "Other Verbs of Putting" classes, all of which undergo similar locative alternations. To these 6, the 8-way task adds the *Run* and Sound Emission verbs, which also undergo locative alternations. The 13-way task includes all of our classes.

The second column of Table 2 includes the accuracy of our supervised learner (the decision tree induction system, C5.0), on the same verb sets as in our clustering experiments. These are the results of a 10-fold cross-validation (with boosting) repeated 50 times.[4] In our earlier work, we found that cross-validation performance averaged about .02, .04, and .11 higher than test performance on the 2-way, 3-way, and multiway tasks, respectively, and so should be taken as an upper bound on what can be achieved.

The third column of Table 2 gives the baseline Acc we calculated from random clusterings. Recall that this is an upper bound on random performance. We use this baseline in calculating reductions in error rate of Acc.

The remaining columns of the table give the Acc, R_{adj}, and $meanSil$ measures as described in Section 4.2, for each of the feature sets we explored in clustering, which we discuss in turn below.

5.1 Full Feature Set

The first subcolumn (Full) under each of the three clustering evaluation measures in Table 2 shows the results using the full set of features (i.e., no feature selection). Although generally higher than the baseline, Acc is well below that of the supervised learner, and R_{adj} and $meanSil$ are generally low.

5.2 Manual Feature Selection

One approach to dimensionality reduction is to hand-select features that one believes to be relevant to a given task. Following Joanis and Stevenson (2003), for each class, we systematically identified the subset of features

[4]These results differ slightly from those reported in Joanis and Stevenson (2003), because of our slight changes in verb sets, discussed in Section 3.2.

Task	C5.0	Base	Acc			R_{adj}			$meanSil$		
		Acc	Full	Ling	Seed	Full	Ling	Seed	Full	Ling	Seed
Benefactive/Recipient	.74	.56	.60	.68	.58	.02	.10	.02	.22	.40	.81
Admire/*Amuse*	.83	.56	.83	.80	.78	.41	.34	.29	.18	.49	.71
Run/Sound Emission	.83	.56	.58	.50	.78	-.00	-.02	.29	.17	.44	.66
Cheat/*Steal–Remove*	.89	.56	.55	.53	.80	-.01	-.02	.34	.30	.29	.74
Wipe/*Steal–Remove*	.78	.56	.65	.73	.70	.07	.18	.15	.24	.33	.89
Mean of 2-way	.81	.56	.64	.65	.73	.10	.12	.22	.22	.39	.76
Spray/*Fill*/Putting	.80	.42	.53	.60	.47	.10	.16	.01	.12	.31	.48
Optionally Intrans.	.66	.42	.38	.38	.58	-.02	-.02	.25	.16	.27	.39
Mean of 3-way	.73	.42	.46	.49	.53	.04	.07	.13	.14	.29	.44
6 Locative Classes	.70	.28	.31	.39	.42	.04	.11	.13	.05	.22	.31
8 Locative Classes	.72	.24	.31	.38	.42	.10	.12	.12	.13	.23	.23
All 13 Classes	.58	.19	.29	.31	.29	.07	.08	.09	.05	.12	.16
Mean of multiway	.67	.23	.30	.36	.38	.07	.10	.11	.08	.19	.23

Table 2: Experimental Results. C5.0 is supervised accuracy; Base Acc is Acc on random clusters. Full is full feature set; Ling is manually selected subset; Seed is seed-verb-selected set. See text for further description.

indicated by the class description given in Levin. For each task, then, the linguistically-relevant subset is defined as the union of these subsets for all the classes in the task.

The results for these feature sets in clustering are given in the second subcolumn (Ling) under each of the Acc, R_{adj}, and $meanSil$ measures in Table 2. On the 2-way tasks, the performance on average is very close to that of the full feature set for the Acc and R_{adj} measures. On the 3-way and multiway tasks, there is a larger performance gain using the subset of features, with an increase in the reduction of the error rate (over Base Acc) of 6-7% over the full feature set.

Overall, there is a small performance gain using the Ling subset of features (with an increase in error rate reduction from 13% to 17%). Moreover, the $meanSil$ value for the manually selected features is almost always very much higher than that of the full feature set, indicating that the subset of features is more focused on the properties that lead to a better separation of the data.

This performance comparison tentatively suggests that good feature selection can be helpful in our task. However, it is important to find a method that does not depend on having an existing classification, since we are interested in applying the approach when such a classification does not exist. In the next two sections, we present unsupervised and minimally supervised approaches to this problem.

5.3 Unsupervised Feature Selection

In order to deal with excessive dimensionality, Dash et al. (1997) propose an unsupervised method to rank a set of features according to their ability to organize the data in space, based on an entropy measure they devise. Unfortunately, this promising method did not prove practical for our data. We performed a number of experiments in which we tested the performance of each feature set from cardinality 1 to the total number of features, where each set of size i differs from the set of size $i-1$ in the addition of the feature with next highest rank (according to the proposed entropy measure). Many feature sets performed very well, and some far outperformed our best results using other feature selection methods. However, across our 10 experimental tasks, there was no consistent range of feature ranks or feature set sizes that was correlated with good performance. While we could have selected a threshold that might work reasonably well with our data, we would have little confidence that it would work well in general, considering the inconsistent pattern of results.

5.4 Semi-Supervised Feature Selection

Unsupervised methods such as Dash et al.'s (1997) are appealing because they require no knowledge external to the data. However, in many aspects of computational linguistics, it has been found that a small amount of labelled data contains sufficient information to allow us to go beyond the limits of completely unsupervised approaches. In our domain in particular, verb class discovery "in a vacuum" is not necessary. A plausible scenario is that researchers would have examples of verbs which they believe fall into different classes of interest, and they want to separate other verbs along the same lines. To model this kind of approach, we selected a sample of five seed verbs from each class. Each set of verbs was judged (by the authors' intuition alone) to be "representative" of the class. We purposely did not carry out any linguistic analysis, although we did check that each verb was reasonably frequent (with log frequencies ranging from 2.6 to 5.1).

For each experimental task, we ran our supervised

Task	Ling	Seed
Benefactive/Recipient	28	5
Admire/*Amuse*	24	4
Run/Sound Emission	21	4
Cheat/*Steal–Remove*	18	4
Wipe/*Steal–Remove*	20	3
Spray/*Fill*/Putting	33	8
Optionally Intrans.	50	10
6 Locative Classes	39	19
8 Locative Classes	46	26
All 13 Classes	72	43

Table 3: Feature counts for Ling and Seed feature sets.

learner (C5.0) on the seed verbs for those classes, in a 5-fold cross-validation (without boosting). We extracted from the resulting decision trees the union of all features used, which formed the reduced feature set for that task. Each clustering experiment used the full set of 20 verbs per class; i.e., seed verbs were included, following our proposed model of guided verb class discovery.[5]

The results using these feature sets are shown in the third subcolumn (Seed) under our three evaluation measures in Table 2. This feature selection method is highly successful, outperforming the full feature set (Full) on Acc and R_{adj} on most tasks, and performing the same or very close on the remainder. Moreover, the seed set of features outperforms the manually selected set (Ling) on over half the tasks. More importantly, the Seed set shows a mean overall reduction in error rate (over Base Acc) of 28%, compared to 17% for the Ling set. The increased reduction in error rate is particularly striking for the 2-way tasks, of 37% for the Seed set compared to 20% for the Ling set.

Another striking result is the difference in $meanSil$ values, which are very much higher than those for Ling (which are in turn much higher than for Full). Thus, not only do we see a sizeable increase in performance, we also obtain tighter and better separated clusters with our proposed feature selection approach.

5.5 Further Discussion

In our clustering experiments, we find that smaller subsets of features generally perform better than the full set of features. (See Table 3 for the number of features in the Ling and Seed sets.) However, not just any small set of features is adequate. We ran 50 experiments using randomly selected sets of features of cardinality $3C$, where C is the number of classes (a simple linear function roughly approximating the number of features in the Seed sets). Mean Acc over these clusterings was much lower than for the Seed sets, and R_{adj} was extremely low (below .08 in all cases). Interestingly, $meanSil$ was generally very high, indicating that there is structure in the data, but not what matches our classification. This confirms that appropriate feature selection, and not just a small number of features, is important for the task of verb class discovery.

We also find that our semi-supervised method (Seed) is linguistically plausible, and performs as well as or better than features manually determined based on linguistic knowledge (Ling). We might also ask, would any subset of verbs do as well? To answer this, we ran experiments using 50 different randomly selected seed verb sets for each class. We found that the mean Acc and $meanSil$ values are the same as that of the Seed set reported above, but mean R_{adj} is a little lower. We tentatively conclude that, yes, any subset of verbs of the appropriate class may be sufficient as a seed set, although some sets are better than others. This is promising for our method, as it shows that the precise selection of a seed set of verbs is not crucial to the success of the semi-supervised approach.

6 Other Verb Clustering Work

Using the same Acc measure as ours, Stevenson and Merlo (1999) achieved performance in clustering very close to that of their supervised classification. However, their study used a small set of five features manually devised for a set of three particular classes. Our feature set is essentially a generalization of theirs, but in scaling up the feature space to be useful across English verb classes in general, we necessarily face a dimensionality problem that did not arise in their research.

Schulte im Walde and Brew (2002) and Schulte im Walde (2003), on the other hand, use a larger set of features intended to be useful for a broad number of classes, as in our work. The R_{adj} scores of Schulte im Walde (2003) range from .09 to .18, while ours range from .02 to .34, with a mean of .17 across all tasks. However, Schulte im Walde's features rely on accurate subcategorization statistics, and her experiments include a much larger set of classes (around 40), each with a much smaller number of verbs (average around 4). Performance differences may be due to the types of features (ours are noisier, but capture information beyond subcat), or due to the number or size of classes. While our R_{adj} results generally decrease with an increase in the number of classes, indicating that our tasks in general may be "easier" than her 40-way distinction, our classes also have many more members (20 versus an average of 4) that need to be grouped together. It is a question for future research to explore the effect of these variables in clustering performance.

[5] We also tried directly applying the mutual information (MI) measure used in decision-tree induction (Quinlan, 1986). We calculated the MI of each feature with respect to the classification of the seed verbs, and computed clusterings using the features above a certain MI threshold. This method did not work as well as running C5.0, which presumably captures important feature interactions that are ignored in the individual MI calculations.

7 Conclusions and Future Work

We have explored manual, unsupervised, and semi-supervised methods for feature selection in a clustering approach for verb class discovery. We find that manual selection of a subset of features based on the known classification performs better than using a full set of noisy features, demonstrating the potential benefit of feature selection in our task. An unsupervised method we tried (Dash et al., 1997) did not prove useful, because of the problem of having no consistent threshold for feature inclusion. We instead proposed a semi-supervised method in which a seed set of verbs is chosen for training a supervised classifier, from which the useful features are extracted for use in clustering. We showed that this feature set outperformed both the full and the manually selected sets of features on all three of our clustering evaluation metrics. Furthermore, the method is relatively insensitive to the precise make-up of the selected seed set.

As successful as our seed set of features is, it still does not achieve the accuracy of a supervised learner. More research is needed on the definition of the general feature space, as well as on the methods for selecting a more useful set of features for clustering. Furthermore, we might question the clustering approach itself, in the context of verb class discovery. Rather than trying to separate a set of new verbs into coherent clusters, we suggest that it may be useful to perform a nearest-neighbour type of classification using a seed set, asking for each new verb "is it like these or not?" In some ways our current clustering task is too easy, because all of the verbs are from one of the target classes. In other ways, however, it is too difficult: the learner has to distinguish multiple classes, rather than focus on the important properties of a single class. Our next step is to explore these issues, and investigate other methods appropriate to the practical problem of grouping verbs in a new language.

Acknowledgements

We are indebted to Allan Jepson for helpful discussions and suggestions. We gratefully acknowledge the financial support of NSERC of Canada and Bell University Labs.

References

Steven Abney. 1991. Parsing by chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*. Kluwer Academic Publ.

M. Dash, H. Liu, and J. Yao. 1997. Dimensionality reduction for unsupervised data. In *Ninth IEEE International Conference on Tools with AI (ICTAI '97)*.

Bonnie J. Dorr and Doug Jones. 1996. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 322–327.

FrameNet. 2003. FrameNet web site. `http://www.icsi.berkeley.edu/~framenet/`.

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

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Eric Joanis and Suzanne Stevenson. 2003. A general feature space for automatic verb classification. In *Proceedings of the Tenth Conference of the EACL (EACL-03)*.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, pages 691–696.

Maria Lapata and Chris Brew. 1999. Using subcategorization to resolve verb class ambiguity. In P. Fung and J. Zhou, editors, *Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pages 266–274. Association for Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago UP.

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

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.

Philip Resnik. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61(1–2):127–159.

Ellen Riloff and Mark Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In *Proceedings of the Sixth WVLC (WVLC-98)*.

Anoop Sarkar and Wootiporn Tripasai. 2002. Learning verb argument structure from minimally annotated corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*.

Sabine Schulte im Walde and Chris Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 223–230.

Sabine Schulte im Walde. 2003. Experiments on the choice of features for learning verb classes. In *Proceedings of the Tenth Conference of the EACL (EACL-03)*.

Suzanne Stevenson and Paola Merlo. 1999. Automatic verb classification using distributions of grammatical features. In *Proceedings of the Ninth Conference of the EACL (EACL-99)*, pages 45–52.

Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. 2000. Impact of similarity measures on web-page clustering. In *Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI-2000)*, pages 58–64.