Improving Verb Clustering with Automatically Acquired Selectional Preferences

Lin Sun and Anna Korhonen
University of Cambridge, Computer Laboratory
15 JJ Thomson Avenue, Cambridge CB3 0GD, UK
ls418,alk23@cl.cam.ac.uk

Abstract

In previous research in automatic verb classification, syntactic features have proved the most useful features, although manual classifications rely heavily on semantic features. We show, in contrast with previous work, that considerable additional improvement can be obtained by using semantic features in automatic classification: verb selectional preferences acquired from corpus data using a fully unsupervised method. We report these promising results using a new framework for verb clustering which incorporates a recent subcategorization acquisition system, rich syntactic-semantic feature sets, and a variation of spectral clustering which performs particularly well in high dimensional feature space.

1 Introduction

Verb classifications have attracted a great deal of interest in natural language processing (NLP). They have proved useful for various important NLP tasks and applications, including e.g. parsing, word sense disambiguation, semantic role labeling, information extraction, question-answering, and machine translation (Swier and Stevenson, 2004; Dang, 2004; Shi and Mihalcea, 2005; Zapirain et al., 2008).

Verb classes are useful because they offer a powerful tool for generalization and abstraction which can be beneficial when faced e.g. with the problem of data sparsity. Particularly useful can be classes which capture generalizations over a range of (cross-)linguistic properties, such as the ones proposed by Levin (1993). Being defined in terms of similar meaning and (morpho-)syntactic behaviour of words, Levin style classes generally incorporate a wider range of properties than e.g. classes defined solely on semantic grounds (Miller, 1995).

In recent years, a variety of approaches have been proposed for automatic induction of verb classes from corpus data (Schulte im Walde, 2006; Joanis et al., 2008; Sun et al., 2008; Li and Brew, 2008; Korhonen et al., 2008; Ó Séaghdha and Copestake, 2008; Vlachos et al., 2009). This work opens up the opportunity of learning and tuning classifications tailored to the application and domain in question. Although manual classification may always yields higher accuracy, automatic verb classification is cost-effective and gathers statistical information as a side-effect of the acquisition process which is difficult for humans to gather but can be highly useful for NLP applications.

To date, both supervised and unsupervised machine learning (ML) methods have been proposed for verb classification and used to classify a variety of features extracted from raw, tagged and/or parsed corpus data. The best performing features on cross-domain verb classification have been syntactic in nature (e.g. syntactic slots, subcategorization frames (SCFs)). Disappointingly, semantic features have not yielded significant additional improvement, although they play a key role in manual and theoretical work on verb classification and could thus be expected to offer a considerable contribution to classification performance.

Since the accuracy of automatic verb classification shows room for improvement, we further investigate the potential of semantic features – verb selectional preferences (SPs) – for the task. We introduce a novel approach to verb clustering which involves the use of (i) a recent subcategorization frame (SCF) acquisition system (Preiss et al., 2007) which produces rich lexical, SCF and syntactic data, (ii) novel syntactic-semantic feature sets extracted from this data which incorporate a variety of linguistic information, including SPs, and (iii) a new variation of spectral cluster-

ing based on the MNCut algorithm (Meila and Shi, 2001) which is well-suited for dealing with the resulting, high dimensional feature space.

Using this approach, we show on two well-established test sets that automatically acquired SPs can be highly useful for verb clustering. They yield high performance when used in combination with syntactic features. We obtain our promising results using a fully unsupervised approach to SP acquisition which differs from previous approaches in that it does not exploit WordNet (Miller, 1995) or other lexical resources. It is based on clustering argument head data in the grammatical relations associated with verbs.

We describe our features in section 2 and the clustering methods in section 3. Experimental evaluation and results are reported in sections 4 and 5, respectively. Section 6 provides discussion and describes related work, and section 7 concludes.

2 Features

Our target classification is the taxonomy of Levin (1993) where verbs taking similar diathesis alternations are assumed to share meaning components and are organized into semantically coherent classes. The main feature of this classification is a diathesis alternation which manifests at the level of syntax in alternating sets of SCF (e.g. in the causative/inchoative alternation an NP frame alternates with an intransitive frame: *Tony broke the window* ↔ *The window broke*).

Since automatic detection of diathesis alternations is very challenging (McCarthy, 2001), most work on automatic classification has exploited the fact that similar alternations tend to result in similar SCFs. The research reported so far[1] has used mainly syntactic features for classification, ranging from shallow syntactic slots (e.g. NPs preceding or following the verb) to SCFs. Some researchers have discovered that supplementing basic syntactic features with information about adjuncts, co-occurrences, tense, and/or voice of the verb have resulted in better performance.

However, additional information about semantic SPs of verbs has not yielded considerable improvement on verb classification although SPs can be strong indicators of diathesis alternations (McCarthy, 2001) and although fairly precise semantic descriptions, including information about verb selectional restrictions, can be assigned to the majority of Levin classes, as demonstrated by VerbNet (Kipper-Schuler, 2005).

SP acquisition from undisambiguated corpus data is arguably challenging (Brockmann and Lapata, 2003; Erk, 2007; Bergsma et al., 2008). It is especially challenging in the context of verb classification where SP models are needed for specific syntactic slots for which the data may be sparse, and the resulting feature vectors integrating both syntactic and semantic features may be high dimensional. However, we wanted to investigate whether better results could be obtained if the features were optimised for richness, the feature extraction for accuracy, and a clustering method capable of dealing with the resulting high dimensional feature space was employed.

2.1 Feature extraction

We adopted a recent SCF acquisition system which has proved more accurate than previous comparable systems[2] but which has not been employed for verb clustering before: the system of Preiss et al. (2007). This system tags, lemmatizes and parses corpus data using the current version of the RASP (Robust Accurate Statistical Parsing) toolkit (Briscoe et al., 2006), and on the basis of resulting grammatical relations (GRs) assigns each occurrence of a verb to one of 168 verbal SCFs classes[3].

The system provides a filter which can be used to remove adjuncts from the resulting lexicon. We do not employ this filter since adjuncts have proved informative for verb classification (Sun et al., 2008; Joanis et al., 2008). However, we do frequency-based thresholding to minimise the noise (e.g. erroneous scfs) and sparse data in verb classification and to ensure that only features supported by several verbs are used in classification: we only consider SCFs and GRs which have frequency larger than 40 with 5 or more verbs[4].

The system produces a rich lexicon which includes raw and processed input sentences and provides a variety of material for verb clustering, including e.g. (statistical) information related to the part-of-speech (POS) tags, GRs, SCFs, argument heads, and adjuncts of verbs. Using this material, we constructed a wide range of feature sets

[1] See section 6 for discussion on previous work.

[2] See Preiss et al. (2007) for the details of evaluation.

[3] We used an implementation of the SCF classifier provided by Paula Buttery.

[4] These and other threshold values mentioned in this paper were determined empirically on corpus data.

for experimentation, both shallow and deep syntactic and semantic features. As described below, some of the feature types have been employed in previous works and some are novel.

2.2 Feature sets

The first feature set F1 includes information about the lexical context (co-occurrences) of verbs which has proved useful for supervised verb classification (Li and Brew, 2008):

F1: Co-occurrence (CO): We adopt the best method of Li and Brew (2008) where collocations are extracted from the four words immediately preceding and following a lemmatized verb. Stop words are removed prior to extraction, and the 600 most frequent resulting COs are kept.

F2-F3 provide information about lexical preferences of verbs in argument head positions of specific GRs associated with the verb:

F2: Prepositional preference (PP): the type and frequency of prepositions in the indirect object relation.

F3: Lexical preference (LP): the type and frequency of nouns and prepositions in the subject, object, and indirect object relation.

All the other feature sets include information about SCFs which have been widely employed in verb classification, e.g. (Schulte im Walde, 2006; Sun et al., 2008; Li and Brew, 2008; Korhonen et al., 2008). F4-F7 include basic SCF information and/or refine it with additional information which has proved useful in previous works:

F4: SCFs and relative frequencies with verbs. SCFs abstract over particles and prepositions.

F5: F4 with COs (F1). The SCF and CO feature vectors are concatenated.

F6: F4 with the tense of the verb. The frequency of verbal POS tags is calculated specific to each SCF.

F7: F4 with PPs (F2). This feature parameterizes SCFs for prepositions.

F8: Basic SCF feature corresponding to F4 but extracted from the VALEX lexicon (Korhonen et al., 2006)[5].

The following 9 feature sets are novel. They build on F7, refining it further. F9-F11 refine F7 with information about LPs:

F9: F7 with F3 (subject only)

F10: F7 with F3 (object only)

F11: F7 with F3 (subject, object, indirect object)

F12-17 refine F7 with SPs. We adopt a fully unsupervised approach to SP acquisition. We acquire the SPs by

1. taking the GR relations (subject, object, indirect object) associated with verbs,

2. extracting all the argument heads in these relations which occur with frequency > 20 with more than 3 verbs, and

3. clustering the resulting N most frequent argument heads into M classes using the spectral clustering method described in the following section.

We tried the N settings $\{200, 500\}$ and the M settings $\{10, 20, 30, 80\}$. The best settings $N = 200, M = 20$ and $N = 500, M = 30$ are reported in this paper. We enforce the features to be shared by all the potential members of a verb class. The expected class size is approximately N/K, and we allow for 10% outliers (the features occurring less than $(N/K) \times 0.9$ verbs are thus removed).

The resulting SPs are combined with SCFs in a similar fashion as LPs are combined with SCFs in F9-F11:

F12-F14: as F9-F11 but SPs (20 clusters from 200 argument heads) are used instead of LPs

F15-F17: as F9-F11 but SPs (30 clusters from 500 argument heads) are used instead of LPs

[5]This feature was included to enable comparing the contribution of the recent SCF system to that of an older, comparable system which was used for constructing the VALEX lexicon.

3 Clustering methods

We use two clustering methods: (i) pairwise clustering (PC) which obtained the best performance in comparison with several other methods in recent work on biomedical verb clustering (Korhonen et al., 2008), and (ii) a method which is new to the task (and to the best of our knowledge, to NLP): a variation of spectral clustering which exploits the MNCut algorithm (Meila and Shi, 2001) (SPEC). Spectral clustering has been shown to be effective for high dimensional and non-convex data in NLP (Chen et al., 2006) and it has been applied to German verb clustering by Brew and Schulte im Walde (2002). However, previous work has used Ng et al. (2002)'s algorithm, while we adopt the MNCut algorithm. The latter has shown a wider applicability (von Luxburg, 2007; Verma and Meila, 2003) and it can be justified from the random walk view, which has a clear probabilistic interpretation.

Clustering groups a given set of items (verbs in our experiment) $V = \{v_n\}_{n=1}^{N}$ into a disjoint partition of K classes $I = \{I_k\}_{k=1}^{K}$. Both our algorithms take a similarity matrix as input. We construct this from the skew divergence (Lee, 2001). The skew divergence between two feature vectors v and v' is $d_{skew}(v, v') = D(v'||a \cdot v + (1-a) \cdot v')$ where D is the KL-divergence. v is smoothed with v'. The level of smoothing is controlled by a whose value is set to a value close to 1 (e.g. 0.9999). We symmetrize the skew divergence as follows: $d(v, v')_{sskew} = \frac{1}{2}(d_{skew}(v, v') + d_{skew}(v', v))$.

SPEC is typically used with the Radial Basis Function (RBF) kernel. We adopt a new kernel similar to the symmetrized KL divergence kernel (Moreno et al., 2004) which avoids the need for scale parameter estimation.

$$w(v, v') = \exp(-d_{sskew}(v, v'))$$

The similarity matrix W is constructed where $W_{ij} = w(v_i, v_j)$.

Pairwise clustering

PC (Puzicha et al., 2000) is a method where a cost criterion guides the search for a suitable partition. This criterion is realized through a cost function of the similarity matrix W and partition I:

$$H = -\sum n_j \cdot \text{Avgsim}_j,$$
$$\text{Avgsim}_j = \frac{\sum_{\{a,b \in A_j\}} w(a,b)}{n_j \cdot (n_j - 1)}$$

where n_j is the size of the j_{th} cluster and Avgsim_j is the average similarity between cluster members.

Spectral clustering

In SPEC, the similarities W_{ij} are viewed as the weight on the edges ij of a graph G over V. The similarity matrix W is thus the adjacency matrix for G. The degree of a vertex i is $d_i = \sum_{j=1}^{N} w_{ij}$. A cut between two partitions A and A' is defined to be $\text{Cut}(A, A') = \sum_{m \in A, n \in A'} W_{mn}$.

In MNCut algorithm, the similarity matrix W is transformed to a stochastic matrix P.

$$P = D^{-1}W \tag{1}$$

The degree matrix D is a diagonal matrix where $D_{ii} = d_i$.

It was shown by Meila and Shi (2001) that if P has the K leading eigenvectors that are piecewise constant[6] with respect to a partition I^* and their eigenvalues are not zero, then I^* minimizes the multiway normalized cut(MNCut):

$$\text{MNCut}(I) = K - \sum_{k=1}^{K} \frac{\text{Cut}(I_k, I_k)}{\text{Cut}(I_k, I)}$$

P_{mn} can be interpreted as the transition probability between vertices m, n. The criterion can thus be expressed as $\text{MNCut}(I) = \sum_{k=1}^{K}(1 - P(I_k \rightarrow I_k|I_k))$ (Meila, 2001), which is the sum of transition probabilities across different clusters. The criterion finds the partition where the random walks are most likely to happen within the same cluster.

In practice, the K leading eigenvectors of P is not piecewise constant. But we can extract the partition by finding the approximately equal elements in the eigenvectors using a clustering algorithm like K-means.

The numerator of MNCut is similar to the cost function of PC. The main differences between the two algorithms are: 1) MNCut takes into account of the cross cluster similarity, while PC does not. 2) PC optimizes the cost function using deterministic annealing, whereas SPEC uses eigensystem decomposition.

The spectral clustering algorithm is based on the Multicut algorithm (Meila and Shi, 2001).

[6] The eigenvector v is piecewise constant with respect to I if $v(i) = v(j) \forall i, j \in I_k$ and $k \in 1, 2...K$

Input: Dataset S, Number of clusters K

1. Compute similarity matrix W and Degree matrix D
2. Construct stochastic matrix P using equation 1
3. Compute the eigenvalues and eigenvectors $\{\lambda_n, x_n\}_{n=1}^{N}$ of P, where $\lambda_n \geq \lambda_{n+1}$, form a matrix $X = [x_2, \ldots, x_k]$ by stacking the eigenvectors in columns.
4. Form a matrix Y from X by normalizing the row sums to have norm 1: $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{\frac{1}{2}}$
5. Consider the row of Y to be the transformed feature vectors for each verb and cluster them into clusters $C_1 \ldots C_k$ using K-means clustering algorithm.

Output: Clusters $C_1 \ldots C_k$

4 Experimental evaluation

4.1 Test sets

We employed two test sets which have been used to evaluate previous work on English verb classification:

T1 The test set of Joanis et al. (2008) provides a classification of 835 verbs into 15 (some coarse, some fine-grained) Levin classes. 11 tests are provided for 2-14 way classifications. We employ the 14 way classification because this corresponds the closest to our target (Levin's fine-grained) classification[7]. We select 586 verbs according to Joanis et al.'s selection criteria, resulting in 10-120 verbs per class. We restrict the class imbalance to 1:1.5.[8]. This yields 205 verbs (10-15 verbs per class) which is similar to the sub-set of T1 employed by Stevenson and Joanis (2003).

T2 The test set of Sun et al. (2008) classifies 204 verbs to 17 fine-grained Levin classes, so that each class has 12 member verbs.

Table 1 shows the classes in T1 and T2.

4.2 Data processing

For each verb in T1 and T2, we extracted all the occurrences (up to 10,000) from the raw corpus data gathered originally for constructing the VALEX lexicon (Korhonen et al., 2006). The data was gathered from five corpora, including e.g. the British National Corpus (Leech, 1992) and the North American News Text Corpus (Graff, 1995). The average frequency of verbs in T1 was 1448 and T2 2166, showing that T1 is a more sparse data set.

The data was first processed using the feature extraction module. Table 2 shows (i) the total number of features in each feature set and (ii) the average per verb in the resulting lexicons for T1 and T2.

We normalized the feature vectors by the sum of the feature values before applying the clustering techniques. Since both clustering algorithms have

[7]However, the correspondence is not perfect with half of the classes including two or more Levin's fine-grained classes.

[8]Otherwise, in the case of a large class imbalance the evaluation measure would be dominated by the classes with large population.

T1		T2	
Object Drop	26.{1,3,7}	Remove	10.1
Recipient	13.{1,3}	Send	11.1
Admire	31.2	Get	13.5.1
Amuse	31.1	Hit	18.1
Run	51.3.2	Amalgamate	22.2
Sound	43.2	Characterize	29.2
Light & Substance	43.{1,4}	Peer	30.3
		Amuse	31.1
Cheat	10.6	Correspond	36.1
Steal & Remove	10.{5,1}	Manner of speaking	37.3
Wipe	10.4.{1,2}	Say	37.7
Spray / Load	9.7	Nonverbal expression	40.2
Fill	9.8		
Putting	9.1-6	Light	43.1
Change of State	45.1-4	Other change of state	45.4
		Mode with Motion	47.3
		Run	51.3.2
		Put	9.1

Table 1: Levin classes in T1 and T2

		T1		T2	
		total	avg	total	avg
CO	F1	1328	764	743	382
LP (p)	F2	61	37	55	25
LP (all)	F3	2521	526	1481	295
SCF	F4	88	46	86	38
SCF+CO	F5	1466	833	856	422
SCF+POS	F6	319	114	299	87
SCF+P	F7	282	96	273	76
SCF (V)	F8	-	-	92	45
SCF+LP (s)	F9	1747	324	1474	225
SCF+LP (o)	F10	2817	424	2319	279
SCF+LP (all)	F11	4250	649	3515	426
SCF+SP20 (s)	F12	821	235	690	145
SCF+SP20 (o)	F13	792	218	706	135
SCF+SP20 (all)	F14	1333	357	1200	231
SCF+SP30 (s)	F15	977	274	903	202
SCF+SP30 (o)	F16	1026	273	1012	205
SCF+SP30 (all)	F17	1720	451	1640	330

Table 2: (i) The total number of features and (ii) the average per verb for all the feature sets

an element of randomness, we run them multiple times. The step 5 of SPEC (K-means) was run for 50 times. The result that minimizes the distortion (the distances to cluster centroid) is reported. PC was run 20 times, and the results are averaged.

4.3 Evaluation measures

To facilitate meaningful comparisons, we employ the same measures for evaluation as previously employed e.g. by Korhonen et al. (2008); Ó Séaghdha and Copestake (2008).

The first measure is modified purity (mPUR) – a global measure which evaluates the mean precision of clusters. Each cluster is associated with its prevalent class. The number of verbs in a cluster K that take this class is denoted by $n_{prevalent}$(K). Verbs that do not take it are considered as errors. Clusters where $n_{prevalent}$(K) = 1 are disregarded as not to introduce a bias towards singletons:

$$\text{mPUR} = \frac{\sum_{n_{prevalent(k_i)>2}} n_{prevalent(k_i)}}{\text{number of verbs}}$$

The second measure is weighted class accuracy (ACC): the proportion of members of dominant clusters DOM-CLUST$_i$ within all classes c_i.

$$\text{ACC} = \frac{\sum_{i=1}^{C} \text{verbs in DOM-CLUST}_i}{\text{number of verbs}}$$

mPUR and ACC can be seen as a measure of precision(P) and recall(R) respectively. We calculate F measure as the harmonic mean of P and R:

$$\text{F} = \frac{2 \cdot \text{mPUR} \cdot \text{ACC}}{\text{mPUR} + \text{ACC}}$$

The random baseline(BL) is calculated as follows:

$$\text{BL} = 1/\text{number of classes}$$

5 Results

5.1 Quantitative evaluation

Table 3 includes the F-measure results for all the feature sets when the two methods (PC and SPEC) are used to cluster verbs in the test sets T1 and T2, respectively. A number of tendencies can be observed in the results. Firstly, the results for T2 are clearly better than those for T1. Including a higher number of verbs lower in frequency from classes of variable granularity, T1 is probably a more challenging test set than T2. T2 is controlled for the number and frequency of verbs to facilitate cross-class comparisons. While this may contribute to better results, T2 is a more accurate test set for us in the sense that it offers a better correspondence with our target (fine-grained Levin) classes.

		T1		T2	
		PC	SPEC	PC	SPEC
	BL	7.14	7.14	5.88	5.88
CO	F1	15.62	33.85	17.86	40.94
LP (p)	F2	40.40	38.97	50.98	49.02
LP (all)	F3	42.94	47.50	41.08	74.55
SCF	F4	34.22	36.16	52.33	57.78
SCF+CO	F5	26.43	28.70	19.52	29.10
SCF+POS	F6	36.14	34.75	44.44	46.70
SCF+P	F7	43.57	43.85	63.40	63.28
SCF (V)	F8	-	-	34.08	38.30
SCF+LP (s)	F9	47.72	56.09	65.94	71.65
SCF+LP (o)	F10	43.09	48.43	57.11	73.97
SCF+LP (all)	F11	45.87	54.63	56.30	72.97
SCF+SP20 (s)	F12	46.67	**57.75**	39.52	71.67
SCF+SP20 (o)	F13	44.95	51.70	40.76	70.78
SCF+SP20(all)	F14	48.19	55.12	39.68	73.09
SCF+SP30 (s)	F15	45.89	56.10	64.44	**80.35**
SCF+SP30 (o)	F16	42.01	48.74	52.75	70.52
SCF+SP30(all)	F17	46.66	52.68	51.07	68.67

Table 3: Results on testsets T1 and T2

Secondly, the difference between the two clustering methods is clear: the new SPEC outperforms PC on both test sets and across all the feature sets. The performance of the two methods is still fairly similar with the more basic, less sparse feature sets (F1-F2, F4, F6-7) but when the more sophisticated feature sets are used (F3, F5, F9-F17) SPEC performs considerably better. This demonstrates that it is clearly a better suited method for high dimensional feature sets.

Comparing the feature sets, the simple co-occurrence based F1 performs clearly better than the random baseline. F2 and F3 which exploit lexical data in the argument head positions of GRs prove significantly better than F1. F3 yields surprisingly good results on T2: it is the second best feature set on this test set. Also on T1, F3 performs better than the SCF-based feature sets F4-F7. This demonstrates the usefulness of lexical data when obtained from argument positions in relevant GRs.

Our basic SCF feature set F4 performs considerably better than the comparable feature set F8 obtained from the VALEX lexicon. The difference is 19.50 in F-measure. As both lexicons were extracted from the same corpus data, the improvement can be attributed to improved parser and SCF acquisition performance (Preiss et al., 2007).

F5-F7 refine the basic SCF feature set F4 further. F5 which combines a SCF with CO information proved the best feature set in the supervised verb classification experiment of Li and Brew (2008). In our experiment, F5 produces substantially lower result than CO and SCF alone (i.e.

F1 and F4). However, our corpus is smaller (Li and Brew used the large Gigaword corpus), our SCFs are different, and our approach is unsupervised, making meaningful comparisons difficult.

F6 combines F4 with information about verb tense. This was not helpful: F6 produces worse results than F4. F7, on the other hand, yields better results than F4 on both test sets. This demonstrates what the previous research has shown: SCF perform better when parameterized for prepositions.

Looking at our novel feature sets F9-F17, F9-F11 combine the most accurate SCF feature set F4 with the LP-based features F2-F3. Although the feature space gets more sparse, all the feature sets outperform F2-F3 on T1. On T2, F3 performs exceptionally well, and thus yields a better result than F9-F11, but F9-F11 nevertheless perform clearly better than the best SCF-based feature set F4 alone. The differences among F9, F10 and F11 are small on T2, but on T1 F9 yields the best performance. It could be that F9 works the best for the more sparse T1 because it suffers the least from data sparsity (it uses LPs only for the subject relation).

F12-F17 replace the LPs in F9-F11 by semantic SPs. When only 20 clusters are used as SP models and acquired from the smaller sample of (200) argument heads (F12-F14), SPs do not perform better than LPs on T2. A small improvement can be observed on T1, especially with F12 which uses only the subject data (yielding the best F measure on T1: 57.75%). However, when 30 more fine-grained clusters are acquired from a bigger sample of (500) argument heads (F15-F17), lower results can be seen on T1. On T2, on the other hand, F15 yields dramatic improvement and we get the best performance for this test set: 80.35% F-measure.

The fact that no improvement is observed when using F16 and F17 on T2 could be explained by the fact that SPs are stronger for the subject position which also suffers less from the sparse data problem than e.g. i. object position. The fact that no improvement is observed on T1 is likely to be due to the fact that verbs have strong SPs only at the finer-grained level of Levin classification. Recall that in T1, as many as half of the classes are coarser-grained.

5.2 Qualitative evaluation

The best performing feature sets on both T1 and T2 were thus our new SP-based feature sets. We conducted qualitative analysis of the best 30 SP clusters in the T2 data created using SPEC to find out whether these clusters were really semantic in nature, i.e. captured semantically meaningful preferences. As no gold standard specific to our verb classification task was available, we did manual cluster analysis using VerbNet (VN) as aid. In VN, Levin classes are assigned with semantic descriptions: the arguments of SCFs involved in diathesis alternations are labeled with thematic roles some of which are labeled with selectional restrictions.

Human	mother, wife, parent, girl, child
Role	patient, student, user, worker, teacher
Body-part	neck, shoulder, back, knee, corner
Authority	committee, police, court, council, board
Organization	society, firm, union, bank, institution
Money	cash, currency, pound, dollar, fund
Amount	proportion, value, size, speed, degree
Time	minute, moment, night, hour, year
Path	street, track, road, stair, route
Building	office, shop, hotel, hospital, house
Region	site, field, area, land, island
Technology	system, model, facility, engine, machine
Task	operation, test, study, analysis, duty
Arrangement	agreement, policy, term, rule, procedure
Matter	aspect, subject, issue, question, case
Problem	difficulty, challenge, loss, pressure, fear
Idea	argument, concept, idea, theory, belief
Power	control, lead, influence, confidence, ability
Form	colour, style, pattern, shape, design
Item	letter, book, goods, flower, card

Table 4: Cluster analysis: 20 clusters, their SP labels, and prototypical member nouns

From the 30 thematic role types in VN, as many as 20 are associated with the 17 Levin classes in T2. The most frequent role in T2 is agent, followed by theme, location, patient, recipient, and source. From the 36 possible selectional restriction types, 7 appear in T2; the most frequent ones being +animate and +organization, followed by +concrete, +location, and +communication.

As SP clusters capture selectional *preferences* rather than *restrictions*, we examined manually whether the 30 clusters (i) capture semantically meaningful classes, and whether they (ii) are plausible given the VN semantic descriptions/restrictions for the classes in T2.

The analysis revealed that all the 30 clusters had a predominant, semantically motivated SP supported by the majority of the member nouns. Although many clusters could be further divided into more specific SPs (and despite the fact that some nouns were clearly misclassified), we were able to assign each cluster a descriptive label characterizing the predominant SP. Table 4 shows 15 sam-

ple clusters, the SP labels assigned to them, and a number of example nouns in these clusters.

When comparing each SP cluster against the VN semantic descriptions/restrictions for T2, we found that each predominant SP was plausible. Also, the SPs frequent in our data were also frequent among the 17 classes according to VN. For example, the many SP clusters labeled as arrangements, issues, ideas and other abstract concepts were also frequent in T2, e.g. among COMMUNICATION (37), CHARACTERISE (29.2), AMALGAMATE (22.2) and other classes.

This analysis showed that the SP models which performed well in verb clustering were semantically meaningful for our task. An independent evaluation using one of the standard datasets available for SP acquisition research (Brockmann and Lapata, 2003) is of course needed to determine how well the acquisition method performs in comparison with other existing methods.

Finally, we evaluated the quality of the verb clusters created using the SP-based features. We found that some of the errors were similar to those seen on T2 when using syntactic features: errors due to polysemy and syntactic idiosyncracy. However, a new error type clearly due to the SP-based feature was detected. A small number of classes got confused because of strong similar SPs in the subject (agent) position. For example, some PEER (30.3) verbs (e.g. *look, peer*) were found in the same cluster with SAY (37.7) verbs (e.g. *shout, yell*) – an error which purely syntactic features do not produce. Such errors were not numerous and could be addressed by developing more balanced SP models across different GRs.

6 Discussion and related work

Although features incorporating semantic information about verb SPs make theoretical sense they have not proved equally promising in previous experiments which have compared them against syntactic features in verb classification. Joanis et al. (2008) incorporated an 'animacy' feature (a kind of a 'SP') which was determined by classifying e.g. pronouns and proper names in data to this single SP class. A small improvement was obtained when this feature was used in conjunction with syntactic features in supervised classification.

Joanis (2002) and Schulte im Walde (2006) experimented with more conventional SPs with syntactic features in English and German verb classification, respectively. They employing top level WordNet (Miller, 1995) and Germanet (Kunze and Lemnitzer, 2002) classes as SP models. Joanis (2002) obtained no improvement over syntactic features, whereas Schulte im Walde (2006) obtained insignificant improvement.

		Method	Result
T1	Li et al. 2008	supervised	66.3
	Joanis et al. 2008	supervised	58.4
	Stevenson et al. 2003	semi-supervised unsupervised	29 31
	SPEC	unsupervised	57.55
T2	Sun et al. 2008	supervised unsupervised	62.50 51.6
	Ó Séaghdha et al. 2008	supervised	67.3
	SPEC	unsupervised	80.35

Table 5: Previous verb classification results

Korhonen et al. (2008) combined SPs with SCFs when clustering biomedical verbs. The SPs were acquired automatically from syntactic slots of SCFs (not from GRs as in our experiment) using PC clustering. A small improvement was obtained using LPs extracted from the same syntactic slots, but the SP clusters offered no improvement. Recently, Schulte im Walde et al. (2008) proposed an interesting SP acquisition method which involves combining EM training and the MDL principle for an verb classification incorporating SPs. However, no comparison against purely syntactic features is provided.

In our experiment, we obtained a considerable improvement over syntactic features, despite using a fully unsupervised approach to both verb clustering and SP acquisition. In addition to the rich, syntactic-semantic feature sets, our good results can be attributed to the clustering technique capable of dealing with them. The potential of spectral clustering for the task was recognised earlier by Brew and Schulte im Walde (2002). Although a different version of the algorithm was employed and applied to German (rather than to English), and although no SP features were used, these earlier experiments did demonstrate the ability of the method to perform well in high dimensional feature space.

To get an idea of how our performance compares with that of related approaches, we examined recent works on verb classification (supervised and unsupervised) which were evaluated on same test sets using comparable evaluation measures. These works are summarized in table 5. ACC and F-measure are shown for T1 and T2, respectively.

On T1, the best performing supervised method reported so far is that of Li and Brew (2008). Li and Brew used Bayesian Multinomial Regression for classification. A range of feature sets integrating COs, SCFs and/or LPs were evaluated. The combination of COs and SCFs gave the best result, shown in the table. Joanis et al. (2008) report the second best supervised result on T1, using Support Vector Machines for classification and features derived from linguistic analysis: syntactic slots, slot overlaps, tense, voice, aspect, and animacy of NPs. Stevenson and Joanis (2003) report a semi- and unsupervised experiment on T1. A feature set similar to that of Joanis et al. (2008) was employed (features were selected in a semi-supervised fashion) and hierarchical clustering was used.

Our unsupervised method SPEC performs substantially better than the unsupervised method of Stevenson et al. and nearly as well as the supervised approach of Joanis et al. (2008) (note, however, that the different experiments involved different sub-sets of T1 so are not entirely comparable).

On T2, the best performing supervised method so far is that of Ó Séaghdha and Copestake (2008) which employs a distributional kernel method to classify SCF features parameterized for prepositions in the automatically acquired VALEX lexicon. Using exactly the same data and feature set, Sun et al. (2008) obtain a slightly lower result when using a supervised method (Gaussian) and a notably lower result when using an unsupervised method (PC clustering). Our method performs considerably better and also outperforms the supervised method of Ó Séaghdha and Copestake (2008).

7 Conclusion and Future Work

We introduced a new approach to verb clustering which involves the use of (i) rich lexical, SCF and GR data produced by a recent SCF system, (ii) novel syntactic-semantic feature sets which combine a variety of linguistic information, and (iii) a new variation of spectral clustering which is particularly suited for dealing with the resulting, high dimensional feature space. Using this approach, we showed on two well-established test sets that automatically acquired SPs can be highly useful for verb clustering. This result contrasts with most previous works but is in line with theoretical work on verb classification which relies not only on syntactic but also on semantic features (Levin, 1993).

In addition to the ideas mentioned earlier, our future plans include looking into optimal ways of acquiring SPs for verb classification. Considerable research has been done on SP acquisition most of which has involved collecting argument headwords from data and generalizing to WordNet classes. Brockmann and Lapata (2003) have showed that WordNet-based approaches do not always outperform simple frequency-based models, and a number of techniques have been recently proposed which may offer ideas for refining our current unsupervised approach (Erk, 2007; Bergsma et al., 2008). The number and type (and combination) of GRs for which SPs can be reliably acquired, especially when the data is sparse, requires also further investigation.

In addition, we plan to investigate other potentially useful features for verb classification (e.g. named entities and preposition classes) and explore semi-automatic ML technology and active learning for guiding the classification. Finally, we plan to conduct a bigger experiment with a larger number of verbs, and conduct evaluation in the context of practical application tasks.

Acknowledgments

Our work was funded by the Dorothy Hodgkin Postgraduate Award, the Royal Society University Research Fellowship, and the EPSRC grant EP/F030061/1, UK. We would like to thank Paula Buttery for letting us use her implementation of the SCF classifier and Yuval Krymolowski for the support he provided for feature extraction.

References

Shane Bergsma, Dekang Lin, and Randy Goebel. Discriminative learning of selectional preference from unlabeled text. In *Proc. of EMNLP*, 2008.

Chris Brew and Sabine Schulte im Walde. Spectral clustering for german verbs. In *Proc. of EMNLP*, 2002.

Ted Briscoe, John Carroll, and Rebecca Watson. The second release of the rasp system. In *Proc. of the COLING/ACL on Interactive presentation sessions*, 2006.

Carsten Brockmann and Mirella Lapata. Evaluating and combining approaches to selectional preference acquisition. In *Proc. of EACL*, 2003.

Jinxiu Chen, Dong-Hong Ji, Chew Lim Tan, and Zheng-Yu Niu. Unsupervised relation disambiguation using spectral clustering. In *Proc. of COLING/ACL*, 2006.

Hoa Trang Dang. *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation*. PhD thesis, CIS, University of Pennsylvania, 2004.

Katrin Erk. A simple, similarity-based model for selectional preferences. In *Proc. of ACL*, 2007.

David Graff. North american news text corpus. *Linguistic Data Consortium*, 1995.

Eric Joanis. Automatic Verb Classification Using a General Feature Space. Master's thesis, University of Toronto, 2002.

Eric Joanis, Suzanne Stevenson, and David James. A general feature space for automatic verb classification. *Natural Language Engineering*, 2008.

Karin Kipper-Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. 2005.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. A large subcategorization lexicon for natural language processing applications. In *Proc. of the 5th LREC*, 2006.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. The Choice of Features for Classification of Verbs in Biomedical Texts. In *Proc. of COLING*, 2008.

Claudia Kunze and Lothar Lemnitzer. GermaNet-representation, visualization, application. In *Proc. of LREC*, 2002.

Lillian. Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*, 2001.

Geoffrey Leech. 100 million words of english: the british national corpus. *Language Research*, 1992.

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

Jianguo Li and Chris Brew. Which Are the Best Features for Automatic Verb Classification. In *Proc. of ACL*, 2008.

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

Marina. Meila. The multicut lemma. Technical report, University of Washington, 2001.

Marina Meila and Jianbo Shi. A random walks view of spectral segmentation. AISTATS, 2001.

George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 1995.

Pedro J. Moreno, Purdy P. Ho, and Nuno Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Proc. of NIPS*, 2004.

Andrew Y. Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proc. of NIPS*, 2002.

Diarmuid Ó Séaghdha and Ann Copestake. Semantic classification with distributional kernels. In *Proc. of COLING*, 2008.

Judita Preiss, Ted Briscoe, and Anna Korhonen. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proc. of ACL*, 2007.

Jan Puzicha, Thomas Hofmann, and Joachim M. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 2000.

Sabine Schulte im Walde. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 2006.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. Combining EM Training and the MDL Principle for an Automatic Verb Classification incorporating Selectional Preferences. In *Proc. of ACL*, pages 496–504, 2008.

Lei Shi and Rada Mihalcea. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Proc. of CICLING*, 2005.

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

Lin Sun, Anna Korhonen, and Yuval Krymolowski. Verb class discovery from rich syntactic data. *Lecture Notes in Computer Science*, 4919:16, 2008.

Robert Swier and Suzanne Stevenson. Unsupervised semantic role labelling. In *Proc. of EMNLP*, 2004.

Deepak Verma and Marina Meila. Comparison of spectral clustering methods. *Advances in Neural Information Processing Systems (NIPS 15)*, 2003.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proc. of the Workshop on Geometrical Models of Natural Language Semantics*, 2009.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 2007.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proc. of ACL*, 2008.