Clustering Polysemic Subcategorization Frame Distributions Semantically

Anna Korhonen
Computer Laboratory
University of Cambridge
15 JJ Thomson Avenue
Cambridge CB3 0FD, UK
Anna.Korhonen@cl.cam.ac.uk

Yuval Krymolowski
Division of Informatics
The University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW,
Scotland UK
ykrymolo@inf.ed.ac.uk

Zvika Marx
The Interdisciplinary Center
for Neural Computation,
The Hebrew University,
Jerusalem, Israel
zvim@cs.huji.ac.il

Abstract

This paper studies unsupervised learning of semantic verb classes through clustering directed by verb subcategorization frames. In contrast to previous work, we provide a line of evidence about the noticeable role polysemy has on the clusters formed. We use the simple nearest neighbours method, as well as an information-theoretic clustering technique, the Information Bottleneck. For evaluation, we introduce principled extensions to standard measures, in order to adapt them to polysemous gold standards. The findings demonstrate that polysemy should be taken into account whenever undisambiguated syntactic data is utilized for revealing semantics.

1 Introduction

Classifications which aim to capture the close relation between the syntax and semantics of verbs have attracted a considerable research interest in both linguistics and computational linguistics (e.g. (Jackendoff, 1990; Levin, 1993; Pinker, 1989; Dang et al., 1998; Dorr, 1997; Merlo and Stevenson, 2001)). While such classifications may not provide a means for full semantic inferencing, they can capture generalizations over a range of linguistic properties, and can therefore be used as a means of reducing redundancy in the lexicon and for filling gaps in lexical knowledge.

Verb classifications have, in fact, been used to support many natural language processing (NLP) tasks, such as language generation, machine translation (Dorr, 1997), document classification (Klavans and Kan, 1998), word sense disambiguation (Dorr and Jones, 1996) and subcategorization acquisition (Korhonen, 2002).

One attractive property of these classifications is that they enable (to a certain extent) inferring the semantics of a verb on the basis of its syntactic behaviour. In recent years several attempts have been made to automatically induce semantic verb classes from (mainly) syntactic information in corpus data (Joanis, 2002; Merlo et al., 2002; Schulte im Walde and Brew, 2002).

In this paper, we focus on the particular task of classifying subcategorization frame (SCF) distributions in a semantically motivated manner. Previous research has demonstrated that clustering can be useful in inferring Levin-style semantic classes (Levin, 1993) from both English and German verb subcategorization information (Brew and Schulte im Walde, 2002; Schulte im Walde, 2000; Schulte im Walde and Brew, 2002).

We propose a novel approach, which involves: (i) obtaining SCF frequency information from a lexicon extracted automatically using the comprehensive system of Briscoe and Carroll (1997) and (ii) applying a clustering mechanism to this information. We use clustering methods that process raw distributional data directly, avoiding complex preprocessing steps required by other advanced methods (e.g. (Brew and Schulte im Walde, 2002)).

In contrast to earlier work, we give special emphasis on polysemy. Earlier work has largely ignored this issue by assuming a single gold standard class for each verb (whether polysemic or not). The relatively good clustering results obtained suggest that many polysemic verbs do have some predominating sense in corpus data. However, this sense can vary across corpora (Roland et al., 2000) and assuming a single sense is inadequate for an important group of medium and high frequency verbs whose distribution of senses in balanced corpus data tends to be flat rather than zipfian (Preiss and Korhonen, 2002).

To investigate the effect of polysemy, we introduce a new evaluation scheme against a polysemic gold standard, which allows for sense variation. This helps to explain the results and offers a better insight into the potential and limits of clustering polysemic SCF data semantically.

We discuss our gold standards and the choice of test verbs in section 2. Section 3 describes the method for subcategorization acquisition and section 4 presents the approach to clustering. Details of the experimental evaluation are supplied in section 5. Section 6 concludes with directions for future work.

2 Semantic Verb Classes and Test Verbs

Levin's taxonomy of verbs and their classes (Levin, 1993) is the largest verb classification in English, employed widely in evaluation of automatic classifications. It provides a classification of 3,024 verbs (4,186 senses) into 48 broad / 192 fine grained classes. Although it is quite extensive, it is not exhaustive. As it primarily concentrates on verbs taking NP and PP complements and does not provide a comprehensive set of senses for verbs, is not suitable for evaluation of polysemic classifications.

We employed as a gold standard a substantially extended version of Levin's classification constructed by Korhonen (2003). This incorporates Levin's classes, 26 additional classes by Dorr (1997)[1], and 57 new classes for verb types not covered comprehensively by Levin or Dorr.

110 test verbs were chosen from this gold standard, 78 polysemic and 32 monosemous ones. Some low frequency verbs were included to investigate the effect of sparse data on clustering performance. To ensure that our gold standard covers all (or most) senses of these verbs, we looked into WordNet (Miller, 1990) and assigned all the WordNet senses of the verbs to gold standard classes[2].

Two versions of the gold standard were created: **monosemous** and **polysemic**. The monosemous one lists only a single sense for each test verb, that corresponding to its predominant (most frequent) sense in WordNet. The polysemic one provides a comprehensive list of senses for each verb. The test verbs and their classes are shown in table 1. The classes are indicated by number codes from the classifications of Levin, Dorr (the classes starting with 0) and Korhonen (the classes starting with A).[3] The predominant sense is indicated by bold font.

3 Subcategorization Information

We obtain our SCF data using the subcategorization acquisition system of Briscoe and Carroll (1997). We expect the use of this system to be beneficial: it employs a robust statistical parser (Briscoe and Carroll, 2002) which yields complete though shallow parses, and a comprehensive SCF classifier, which incorporates 163 SCF distinctions, a superset of those found in the ANLT (Boguraev et al., 1987) and COMLEX (Grishman et al., 1994) dictionaries. The SCFs abstract over specific lexically-governed particles and prepositions and specific predicate selectional preferences but include some derived semi-predictable bounded dependency constructions, such as particle and dative movement.

78 of these 'coarse-grained' SCFs appeared in our data. In addition, a set of 160 fine grained frames were employed. These were obtained by parameterizing two high frequency SCFs for prepositions: the simple PP and NP + PP frames. The scope was restricted to these two frames to prevent sparse data problems in clustering.

A SCF lexicon was acquired using this system from the British National Corpus (Leech, 1992, BNC) so that a maximum of 7000 citations were

[1] These classes are incorporated in the 'LCS database' (http://www.umiacs.umd.edu/∼bonnie/verbs-English.lcs).

[2] As WordNet incorporates particularly fine grained sense distinctions, some senses were found which did not appear in our gold standard. As many of them appeared marginal and/or low in frequency, we did not consider these additional senses in our experiment.

[3] The gold standard assumes Levin's broad classes (e.g. class 10) instead of possible fine-grained ones (e.g. class 10.1).

TEST VERB	GOLD STANDARD CLASSES	TEST VERB	GOLD STANDARD CLASSES	TEST VERB	GOLD STANDARD CLASSES	TEST VERB	GOLD STANDARD CLASSES
place	**9**	*colour*	**24**, 31, 45	*concentrate*	**31**, 45	*look*	**30**, 35
lay	**9**	*dye*	**24**, 21, 41	*focus*	**31**, 45	*stare*	**30**
drop	**9**, 45, 004, 47, 51, A54, A30	*build*	**26**, 45	*force*	**002**, 11	*glow*	**43**
pour	**9**, 43, 26, 57, 13, 31	*bake*	**26**, 45	*persuade*	**002**	*sparkle*	**43**
load	**9**	*invent*	**26**, 27	*urge*	**002**, 37	*dry*	**45**
settle	**9**, 46, A16, 36, 55	*publish*	**26**, 25	*want*	**002**, 005, 29, 32	*shut*	**45**
fill	**9**, 45, 47	*cause*	**27**, 002	*need*	**002**, 005, 29, 32	*hang*	**47**, 9, 42, 40
remove	**10**, 11, 42	*generate*	**27**, 13, 26	*grasp*	**30**, 15	*sit*	**47**, 9
withdraw	**10**, A30	*induce*	**27**, 002, 26	*understand*	**30**	*disappear*	**48**
wipe	**10**, 9	*acknowledge*	**29**, A25, A35	*conceive*	**30**, 29, A56	*vanish*	**48**
brush	**10**, 9, 41, 18	*proclaim*	**29**, 37, A25	*consider*	**30**, 29	*march*	**51**
filter	**10**	*remember*	**29**, 30	*perceive*	**30**	*walk*	**51**
send	**11**, A55	*imagine*	**29**, 30	*analyse*	**34**, 35	*travel*	**51**
ship	**11**, A58	*specify*	**29**	*evaluate*	**34**, 35	*hurry*	**53**, 51
transport	**11**, 31	*establish*	**29**, A56	*explore*	**35**, 34	*rush*	**53**, 51
carry	**11**, 54	*suppose*	**29**, 37	*investigate*	**35**, 34	*begin*	**55**
drag	**11**, 35, 51, 002	*assume*	**29**, A35, A57	*agree*	**36**, 22, A42	*continue*	**55**, 47, 51
push	**11**, 12, 23, 9, 002	*think*	**29**, 005	*communicate*	**36**, 11	*snow*	**57**, 002
pull	**11**, 12, 13, 23, 40, 016	*confirm*	**29**	*shout*	**37**	*rain*	**57**
give	**13**	*believe*	**29**, 31, 33	*whisper*	**37**	*sin*	**003**
lend	**13**	*admit*	**29**, 024, 045, 37	*talk*	**37**	*rebel*	**003**
study	**14**, 30, 34, 35	*allow*	**29**, 024, 13, 002	*speak*	**37**	*risk*	**008**, A7
hit	**18**, 17, 47, A56, 31, 42	*act*	**29**	*say*	**37**, 002	*gamble*	**008**, 009
bang	**18**, 43, 9, 47, 36	*behave*	**29**	*mention*	**37**	*beg*	**015**, 32
carve	**21**, 25, 26	*feel*	**30**, 31, 35, 29	*eat*	**39**	*pray*	**015**, 32
add	**22**, 37, A56	*see*	**30**, 29	*drink*	**39**	*seem*	**020**
mix	**22**, 26, 36	*hear*	**30**, A32	*laugh*	**40**, 37	*appear*	**020**, 48, 29
colour	**24**, 31, 45	*notice*	**30**, A32	*smile*	**40**, 37		

Table 1: Test verbs and their monosemous/polysemic gold standard

used per test verb. The lexicon was evaluated against manually analysed corpus data after an empirically defined threshold of 0.025 was set on relative frequencies of SCFs to remove noisy SCFs out. The method yielded 71.8% precision and 34.5% recall. When we removed the filtering threshold, and evaluated the noisy distribution, F-measure[4] dropped from 44.9 to 38.51.[5]

4 Clustering Method

Data clustering is a process which aims to partition a given set into subsets (clusters) of elements that are similar to one another, while ensuring that elements that are not similar are assigned to different clusters. We use clustering for partitioning a set of verbs. Our hypothesis is that information about SCFs and their associated frequencies is relevant for identifying semantically related verbs. Hence, we use SCFs as *relevance features* to guide the clustering process.[6]

In our work, we try to avoid task-oriented tuning, such as pre-fixed thresholds or restricted cluster sizes, used in some earlier verb clustering works. Recently, a more principled technique has been applied by Brew and Schulte im Walde (2002) which involves performing spectral decomposition and feature selection prior to clustering. While these approaches are worth investigating, we believe that along with noise filtered using such pre-processing steps, valuable information might be lost as well. We prefer methods which approach data more straightforwardly, in its raw distributional form.

We use two methods: (i) a simple hard clustering method that collects the nearest neighbours (NN) of each verb (figure 1), and (ii) the Information Bottleneck (IB), an iterative soft clustering method (Tishby et al., 1999) based on information-theoretic grounds.

The NN method is very simple, but has some disadvantages. It deterministically outputs only one clustering configuration, not allowing the examination of different cluster granularities. It is also highly sensitive to noise: few exceptional neighbourhood relations contradicting the typical trends in the data would be enough to cause the formation of a single cluster which encompasses all elements. Therefore,

[4] $F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

[5] These figures are not particularly impressive because our evaluation is exceptionally hard. We use 1) highly polysemic test verbs, 2) a high number of SCFs and 3) evaluate against manually analysed data rather than dictionaries (the latter have high precision but low recall).

[6] The relevance of the features to the task is evident when looking at the probability of a random $verb_i$ to share the same predominant sense with another randomly chosen $verb_j$ (4.5%) and when comparing this to the probability obtained when $verb_j$ is the JS-divergence nearest neighbor of $verb_i$ (36%) (see figure 1 for the definition of NN).

```
**NN Clustering:**
  1. For each verb $v$:
  2. Calculate the JS divergence between the SCF
     distributions of $v$ and all other verbs:
       $\text{JS}(p,q) = \frac{1}{2}\left[D\left(p\middle\|\frac{p+q}{2}\right) + D\left(q\middle\|\frac{p+q}{2}\right)\right]$
  3. Connect $v$ with the most similar verb;
  4. Find all the connected components
```

Figure 1: Connected components nearest neighbour (NN) clustering. D is the Kullback-Leibler distance.

although the NN method produced interesting results (see section 5), we employed the more sophisticated IB method as well.

The IB method approaches data clustering from an information-theoretic perspective. It quantifies the *relevance information* of a SCF distribution with respect to output clusters, through their mutual information $I(Clusters; \text{SCF}s)$. The relevance information is maximized, while the *compression information* $I(Clusters; Verbs)$ is minimized. This ensures optimal compression of data through clusters. The tradeoff between the two constraints is realized through minimizing the cost term:

$$\mathcal{L} = I(Clusters; Verbs) - \beta I(Clusters; \text{SCF}s)\,,$$

where β is a parameter that balances the constraints.

The IB iterative algorithm finds a local minimum of the above cost term. It takes three inputs: (i) SCF-verb distributions, (ii) the desired number of clusters $\mathcal{K}$, and (iii) the value of β. (For a certain $\mathcal{K}$, there is some minimum possible β that increases with $\mathcal{K}$. An external loop modifies β until this value is reached).

Starting from a random configuration, the algorithm repeatedly calculates, for each cluster K verb V and SCF S, the following probabilities: (i) the marginal proportion of the cluster $p(K)$; (ii) the probability $p(S|K)$ for a SCF to occur with members of the cluster; and (iii) the probability $p(K|V)$ for a verb to be assigned to the cluster. These probabilities are used, each in its turn, for calculating the other probabilities (figure 2). The collection of all $p(S|K)$'s for a fixed cluster K can be regarded as a probabilistic center (*centroid*) of that cluster in the SCF space.

The IB algorithm delivers as output the assignment probabilities $p(K|V)$. Although we could use these probabilities for "soft" clustering (e.g. assign a verb V to several clusters instead of just one) we currently "harden" the output and assign each V to the most probable cluster $K(V)$ only[7]:

```
**IB Clustering** (fixed $\beta$):
Perform till convergence, for each time step
$t = 1, 2, \ldots$ :
  1. $z_t(K,V) = p_{t-1}(K)\, e^{-\beta D[p(S|V) \| p_{t-1}(S|K)]}$
     (When $t = 1$, initialize $z_t(K,V)$ arbitrarily)
  2. $p_t(K|V) = \frac{z_t(K,V)}{\sum_{K'} z_t(K',V)}$
  3. $p_t(K) = \sum_V p(V) p_t(K|V)$
  4. $p_t(S|K) = \sum_V p(S|V) p_t(V|K)$
```

Figure 2: Information Bottleneck (IB) iterative clustering. D is the Kullback-Leibler distance.

$$K(V) = \operatorname*{argmax}_K p(K|V)\,.$$

The IB method gives an indication for the most informative output configurations.[8] It turns out that intensifying the weight on the relevance information $I(Clusters; \text{SCF}s)$, i.e. introducing in repeated runs gradually incremented β values to the IB iterative algorithm (figure 2), allows the production of a larger number of clusters (with too small β, some of the clusters obtained are identical to one another). The relevance information grows when $\mathcal{K}$ and β increase. Those output configurations are regarded as informative where the relevance increases more sharply between $\mathcal{K} - 1$ and $\mathcal{K}$ clusters, than between $\mathcal{K}$ and $\mathcal{K} + 1$ to $\mathcal{K} + 2$.

5 Experimental Evaluation

5.1 Method

A number of different strategies have been proposed for evaluation of clustering. While there is little theoretical consensus on the best strategy, it is clear that the choice of a method should ultimately depend on the task at hand.

[7] This approach was taken to simplify our experiments. However, we will study soft clustering in the future as it offers a means to both investigate and address polysemy.

[8] Most works on clustering ignore this issue and refer to an arbitrarily chosen number of clusters, or to the number of gold standard classes, which cannot be assumed in realistic applications.

Here, we experiment with some measures which have previously been used for evaluating hard verb clusters against a monosemous gold standard. As we currently assign a single sense to each polysemic verb (sec. 5.3) these measures are also applicable for evaluation against a polysemous gold standard.

The first method – the pairwise approach – evaluates clusters in terms of verb pairs. Although it is somewhat questionable (strictly speaking, we do not use clustering to classify verbs into pairs), we include it here for comparison.

We measure the *non-weighted averaged pairwise precision*, i.e., the average proportion of all within-cluster pairs that are correctly co-assigned:

$$\mathrm{AvPP} = \frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} \frac{\text{number of correct pairs in } k_i}{\text{number of pairs in } k_i}$$

where $\{k_i\}_{i=1}^{\mathcal{K}}$ denotes the set of clusters.

The AvPP scores clusters irrespective of their size. In order to favour larger clusters over small ones we use the *adjusted pairwise precision* (APP):

$$\mathrm{APP} = \frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} \frac{\text{num. of correct pairs in } k_i}{\text{num. of pairs in } k_i} \cdot \frac{|k_i|-1}{|k_i|+1} \,.$$

The APP penalizes clusters by a factor that increases with cluster size.[9]

The second approach we adopt bases evaluation on the matrix $h(k,c)$ of cluster-class overlap counts. This approach is more suitable for our task than the pairwise approach, as it evaluates the acquired classes as integral entities rather than collections of pairs. If one thinks of the target as cluster-class one-to-one exclusive match, then only the dominant, most prevalent semantic class within each cluster should be considered.

We denote the number of verbs in a cluster K that take its prevalent class by $n_{\text{prevalent}}(K)$. Verbs that do not take this class are considered as errors. Using $n_{\text{prevalent}}$, we define the *weighted cluster purity*: the proportion of verbs in dominant classes among all clustered verbs.

$$\mathrm{PUR} = \frac{\sum_{i=1}^{\mathcal{K}} n_{\text{prevalent}}(k_i)}{\text{number of verbs}}$$

This measure favours a large number of clusters. A perfect result is obtained for the trivial configuration where each cluster contains a single element. To cover for this bias, we introduce the *modified purity*, where we accept only those dominant classes which include two or more verbs:

$$m\mathrm{PUR} = \frac{\sum_{n_{\text{prevalent}}(k_i) \geq 2} n_{\text{prevalent}}(k_i)}{\text{number of verbs}} \,.$$

Alg.	$\mathcal{K}$	+PP	–PP	+PP	–PP
		AvPP:		APP:	
NN	(24)	45%	41%	20%	18%
IB	42	35%	20%	15%	9%
		PUR:		mPUR:	
NN	(24)	50%	47%	48%	45%
IB	42	66%	60%	50%	38%

Table 2: Clustering performance on the predominant senses, with and without prepositions

5.2 Evaluation Against the Predominant Sense

We first evaluated the clusters against the monosemous gold standard, assuming SCFs with and without prepositions (section 2).

The NN algorithm produced 24 clusters. We requested from the IB algorithm $\mathcal{K} = 2$ to 60 clusters on the 110-verb input. The upper limit was chosen so as to slightly exceed the case when the average cluster size $110/\mathcal{K} = 2$. The counts for these runs were extracted from unfiltered (noisy) SCF distributions.[10]

We report the IB results for $\mathcal{K} = 25,\ 35$ and 42. For these values, the SCF relevance satisfies our criterion for a notable improvement in cluster quality (see the end of section 4). The value $\mathcal{K} = 35$ is very close to the actual number (34) of predominant senses in the gold standard. In this way, the IB yields structural information beyond clustering.

The results against the predominant sense are shown in Table 2. The benefit of using fine-grained SCFs (with prepositions) is evident. The fact that these results are fairly good (comparable to e.g. Schulte im Walde and Brew (2002)) indicates that most of our test verbs indeed have a single predominating sense in balanced corpus data. We ar-

[9] Our definition differs by a factor of 2 from that of Schulte im Walde and Brew (2002).

[10] This yielded better results, which might indicate that the unfiltered "noisy" SCFs contain information which is valuable for the task.

Different Senses	Pairs	Fraction in cluster
0	39	51%
1	85	10%
2	625	7%
3	1284	3%
4	1437	3%

Table 3: The fraction of verb pairs clustered together, as a function of the number of different senses between pair members (results of the NN algorithm)

Common	one irregular		no irregular	
Senses	Pairs	in cluster	Pairs	in cluster
0	2180	3%	3018	3%
1	388	9%	331	12%
2	44	20%	31	35%

Table 4: The fraction of verb pairs clustered together, as a function of the number of shared senses (results of the NN algorithm)

$\mathcal{K}$	Pred. sense	Multiple senses	Pred. sense	Multiple senses
	AvPP:		APP:	
NN: (24)	46%	60% (49% + 4σ)	21%	29% (23%+ 5σ)
IB: 25	24%	35% (27% + 5σ)	12%	18% (14%+ 5σ)
35	32%	43% (34% + 5σ)	14%	20% (16%+ 6σ)
42	35%	46% (38% + 4σ)	15%	19% (16%+ 3σ)
	PUR:		mPUR:	
NN: (24)	50%	60% (54% + 3σ)	48%	60% (46%+ 2σ)
IB: 25	46%	54% (50% + 3σ)	39%	48% (43%+ 3σ)
35	58%	69% (61% + 6σ)	47%	59% (50%+ 4σ)
42	66%	72% (69% + 2σ)	50%	59% (54%+ 2σ)

Table 5: Evaluation against the monosemous (Pred.) and polysemous (Multiple) gold standards. The figures in parentheses are results of evaluation on randomly polysemous data + significance of the actual figure. Results were obtained with fine-grained SCFs (including prepositions).

gue, however, that evaluation against a monosemous gold standard reveals only part of the picture.

5.3 Evaluation Against Multiple Senses

In evaluation against the polysemic gold standard, we assume that a verb polysemous in our corpus data may appear in a cluster with verbs that share any of its senses. In order to evaluate the clusters against polysemous data, we assigned each polysemic verb V a single sense: the one it shares with the highest number of verbs in the cluster $K(V)$.

Tables 3 and 4 show that polysemy has a direct impact on clusters. Table 3 demonstrates that the more two verbs differ by their senses, the lower their chance of ending up in the same cluster. From table 4 we see that the probability of two verbs to appear in the same cluster (NN results) increases with the number of senses they share. However, it is not only the *degree* of polysemy which influcences the results, but also the *type*. For verb pairs where at least one of the members displays ‘irregular’ polysemy (i.e. it does not share its *full* set of senses with any other verb), the probability of co-occurrence in the same cluster is far lower than for verbs which are polysemic in a ‘regular’ manner (Table 4).

In order to show that polysemy makes a non-trivial contribution in shaping the clusters, we measured the improvement that can be due to pure chance by creating randomly polysemous gold standards. We constructed 100 sets of random gold standards. In each iteration, the verbs kept their original predominant senses, but the set of additional senses was taken entirely from another verb - chosen at random. By doing so, we preserved the dominant sense of each verb, the total frequency of all senses and the correlations between the additional senses.

The results, presented in table 5, indicate that the improvement is significant at 3σ level (except three cases, where the significance is 2σ), i.e., the probability of it being an artificial by-product of polysemy is less than 0.5%.

5.4 Qualitative Analysis of the Results

We also performed qualitative analysis by hand to further investigate the effect of polysemy. Consider the following representative clusters:

A1: *talk* (37), *speak* (37)
A2: *look* (30, 35), *stare* (30)
A3: *focus* (31, 45), *concentrate* (31, 45)
A4: *add* (22.1, 37.7, A56)

We observed close relation between the clustering performace and the following patterns of semantic behaviour:

1) Monosemy: We had 32 monosemous test verbs. 10 gold standard classes included 2 or more or these. 7 classes were correctly acquired using clustering (e.g. **A1**), indicating that clustering monosemous verbs is fairly ‘easy’.

2) Predominant sense: 10 clusters were examined by hand whose members got correctly classified together, despite one of them being polysemous (e.g. **A2**). In 8 cases there was a clear indication in the data (when examining SCFs and the selectional preferences on argument heads) that the polysemous verb indeed had its predominant sense in the relevant class and that the co-occurrence was not due to noise.

3) Regular Polysemy: Several clusters were produced which represent linguistically plausible intersective classes (e.g. **A3**) (Dang et al., 1998) rather than single classes.

4) Irregular Polysemy: Verbs with irregular polysemy[11] were frequently assigned to singleton clusters. For example, *add* (**A4**) has a ‘combining and attaching’ sense in class 22 which involves NP and PP SCFs and another ‘communication’ sense in 37 which takes sentential SCFs. Irregular polysemy was not a marginal phenomenon: it explains 5 of the 10 singletons in the IB output with $\mathcal{K} = 42$.

Finally, we performed a qualitative analysis of errors. Consider the following clusters:

B1: *place* (9), *build* (26, 26, 45),
publish (26, 25), *carve* (21, 25, 26)

B2: *sin* (003), *rain* (57), *snow* (57, 002)

B3: *agree* (36, 22, A42), *appear* (020, 48, 29),
begin (55), *continue* (55, 47, 51)

B4: *beg* (015, 32)

The following error types were identified:

1) Syntactic idiosyncracy: This was the most frequent error type, exemplified in **B1**, where *place* is incorrectly clustered with *build*, *publish* and *carve* merely because it takes prepositions similar than these verbs (e.g. *in, on, into*).

2) Sparse data: Many of the low frequency verbs (we had 12 with frequency less than 300) performed poorly. In **B2**, *sin* (which had 53 occurrences) is classified with *rain* and *snow* because it does not occur in our data with the preposition *against* - the ‘hallmark’ of its gold standard class (’Conspire Verbs’).

3) Problems in SCF acquisition: These were not numerous but occurred e.g. when the system could not distinguish between different control (e.g. subject/object equi/raising) constructions (**B3**).

4) Gaps in the gold standard classification: **B4** shows that *beg* ended up in a singleton cluster, despite sharing its both gold standard classes with *pray*. The problem is that *pray* occurs frequently in another sense, ‘address to God’, which gives rise to an intransitive SCF (e.g. *he prayed all the day long*). This sense is too infrequent among verbs to construct a meaningful class for it.

6 Discussion and Conclusions

This paper presented a new approach to automatic semantic classification of verbs. It involved applying the Information Bottleneck and NN methods to cluster polysemic SCF distributions extracted from corpus data using Briscoe and Carroll’s (1997) system. A principled evaluation procedure was performed which allowed to investigate the effect of polysemy on the resulting classification.

Our investigation revealed that polysemy has a considerable role on the clusters formed: polysemic verbs with a clear predominant sense and those which show similar regular polysemy get frequently classified together. Homonymic verbs or verbs with strong irregular polysemy tend to resist any classification.

While we believe that evaluation should account for these cases rather than ignore them, it is clear that the issue of polysemy is related to another, bigger issue: the potential and limitations of clustering in inducing semantic information from polysemic SCF data. Our results show that it is unrealistic to expect that the most ‘important’ (high frequency) verbs in language fall into classes corresponding to single senses. However, our investigation also suggests that clustering can be used for novel, previously unexplored purposes: to detect from corpus data general patterns of semantic behaviour (monosemy, predominant sense, regular/irregular polysemy).

Future work will involve improving the accuracy of subcategorization acquisition, investigating the role of noise (irregular / regular) in clustering, ex-

[11] Recall our definition of irregular polysemy, section 5.3.

amining whether different syntactic/semantic verb types require different approach in clustering, developing our gold standard classification further, and extending our experiments to a larger number of verbs and verb classes. In addition, we plan to investigate the use of soft clustering (without hardening the output) and develop methods for evaluating the soft output against polysemous gold standards.

References

B. Boguraev, E. J. Briscoe, J. Carroll, D. Carter, and C. Grover. 1987. The derivation of a grammatically-indexed lexicon from the longman dictionary of contemporary english. In *Proc. of the* 25th *ACL*, pages 193–200, Stanford, CA.

C. Brew and S. Schulte im Walde. 2002. Spectral clustering for german verbs. In *Conference on Empirical Methods in Natural Language Processing*, Philadelphia, USA.

E. J. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In 5th *ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington DC.

E. J. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In 3rd *International Conference on Language Resources and Evaluation*, pages 1499–1504, Las Palmas, Gran Canaria.

H. T. Dang, K. Kipper, M. Palmer, and J. Rosenzweig. 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proc. of COLING/ACL*, pages 293–299, Montreal, Canada.

B. Dorr and D. Jones. 1996. Role of word sense disambiguation in lexical acquisition: predicting semantics from syntactic cues. In 16th *International Conference on Computational Linguistics*, pages 322–333, Copenhagen, Denmark.

B. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–325.

R. Grishman, C. Macleod, and A. Meyers. 1994. Comlex syntax: building a computational lexicon. In *International Conference on Computational Linguistics*, pages 268–272, Kyoto, Japan.

R. Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, Massachusetts.

E. Joanis. 2002. Automatic verb classification using a general feature space. Master's thesis, University of Toronto.

J. L. Klavans and M. Kan. 1998. Role of verbs in document analysis. In *Proc. of COLING/ACL*, pages 680–686, Montreal, Canada.

A. Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, University of Cambridge, UK.

A. Korhonen. 2003. *Extending Levin's Classification with New Verb Classes*. Unpublished manuscript, University of Cambridge Computer Laboratory.

G. Leech. 1992. 100 million words of english: the british national corpus. *Language Research*, 28(1):1–13.

B. Levin. 1993. *English Verb Classes and Alternations*. Chicago University Press, Chicago.

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

P. Merlo, S. Stevenson, V. Tsang, and G. Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proc. of the* 40th *ACL*, Pennsylvania, USA.

G. A. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

S. Pinker. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, Massachusetts.

J. Preiss and A. Korhonen. 2002. Improving subcategorization acquisition with wsd. In *ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, USA.

D. Roland, D. Jurafsky, L. Menn, S. Gahl, E. Elder, and C. Riddoch. 2000. Verb subcatecorization frequency differences between business-news and balanced corpora. In *ACL Workshop on Comparing Corpora*, pages 28–34.

S. Schulte im Walde and C. Brew. 2002. Inducing german semantic verb classes from purely syntactic subcategorisation information. In *Proc. of the* 40th *ACL*, Philadephia, USA.

S. Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proc. of COLING-2000*, pages 747–753, Saarbrücken, Germany.

N. Tishby, F. C. Pereira, and W. Bialek. 1999. The information bottleneck method. In *Proc. of the* 37th *Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.