The Choice of Features for Classification of Verbs in Biomedical Texts

Anna Korhonen
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
Cambridge CB3 0FD, UK
alk23@cl.cam.ac.uk

Yuval Krymolowski
Dept. of Computer Science
Haifa University
Israel
yuvalkry@gmail.com

Nigel Collier
National Institute of Informatics
Hitotsubashi 2-1-2
Chiyoda-ku, Tokyo 101-8430
Japan
collier@nii.ac.jp

Abstract

We conduct large-scale experiments to investigate optimal features for classification of verbs in biomedical texts. We introduce a range of feature sets and associated extraction techniques, and evaluate them thoroughly using a robust method new to the task: cost-based framework for pairwise clustering. Our best results compare favourably with earlier ones. Interestingly, they are obtained with sophisticated feature sets which include lexical and semantic information about selectional preferences of verbs. The latter are acquired automatically from corpus data using a fully unsupervised method.

1 Introduction

Recent years have seen a massive growth in the scientific literature in the domain of biomedicine. Because future research in the biomedical sciences depends on making use of all this existing knowledge, there is a strong need for the development of natural language processing (NLP) tools which can be used to automatically locate, organize and manage facts related to published experimental results.

Major progress has been made on information retrieval and on the extraction of specific relations (e.g. between proteins and cell types) from biomedical texts (Ananiadou et al., 2006). Other tasks, such as the extraction of factual information, remain a bigger challenge.

Researchers have recently begun to use deeper NLP techniques (e.g. statistical parsing) for improved processing of the challenging linguistic structures (e.g. complex nominals, modal subordination, anaphoric links) in biomedical texts. For optimal performance, many of these techniques require richer syntactic and semantic information than is provided by existing domain lexicons (e.g. UMLS metathesaurus and lexicon[1]). This particularly applies to verbs, which are central to the structure and meaning of sentences.

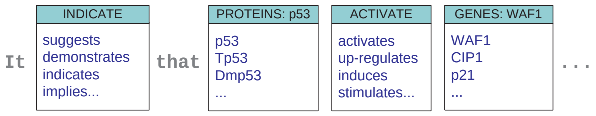

Figure 1: Sample lexical classes

Where the information is absent, *lexical classification* can compensate for it, or aid in obtaining it. Lexical classes which capture the close relation between the syntax and semantics of verbs provide generalizations about a range of linguistic properties (Levin, 1993). For example, consider the INDICATE and ACTIVATE verb classes in Figure 1. Their members have similar subcategorization frames SCFs (e.g. *activate / up-regulate / induce / stimulate* NP) and selectional preferences (e.g. *activate / up-regulate / induce / stimulate* GENES:WAF1), and they can be used to make similar statements describing similar events (e.g. PROTEINS:P53 ACTIVATE GENES:WAF1).

Lexical classes can be used to abstract away from individual words, or to build a lexical organization which predicts much of the behaviour of a new word by associating it with an appropriate class. They have proved useful for various NLP application tasks, e.g. parsing, word sense dis-

[1] http://www.nlm.nih.gov/research/umls

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 449–456
Manchester, August 2008

ambiguation, semantic role labeling, information extraction, question-answering, machine translation (Dorr, 1997; Prescher et al., 2000; Swier and Stevenson, 2004; Dang, 2004; Shi and Mihalcea, 2005). A large-scale classification specific to the biomedical data could support key BIO-NLP tasks such as anaphora resolution, predicate-argument identification, event extraction and the identification of biomedical (e.g. interaction) relations. However, no such classification is available.

Recent research shows that it is possible to automatically induce lexical classes from corpora with promising accuracy (Schulte im Walde, 2006; Joanis et al., 2007; Sun et al., 2008). A number of machine learning (ML) methods have been applied to classify mainly syntactic features (e.g. subcategorization frames (SCFs)) extracted from cross-domain corpora using e.g. part-of-speech tagging or robust statistical parsing techniques. Korhonen et al. (2006) have recently applied such an approach to biomedical texts. Their preliminary experiment shows encouraging results but further work is required before such an approach can be used to benefit practical BIO-NLP.

We conduct a large-scale investigation to find optimal features for biomedical verb classification. We introduce a range of theoretically-motivated feature sets and evaluate them thoroughly using a robust method new to the task: a cost-based framework for pairwise clustering. Our best results compare favourably with earlier ones. Interestingly, they are obtained using feature sets which have proved challenging in general language verb classification: ones which incorporate information about selectional preferences of verbs. Unlike in earlier work, we acquire the latter from corpus data using a fully unsupervised method.

We present our lexical classification approach in section 2 and data in section 3. Experimental evaluation is reported in section 4. Section 5 provides discussion and section 6 concludes.

2 Approach

Our lexical classification approach involves (i) extracting features from corpus data and (ii) clustering them. These steps are described in the following two sections, respectively.

2.1 Features

Lexical classifications are based on diathesis alternations which manifest in alternating sets of syntactic frames (Levin, 1993). Most verb classification approaches have therefore employed shallow syntactic slots or SCFs as basic features. Some have supplemented them with further information about verb tense, voice, and/or semantic selectional preferences on argument heads.[2]

The preliminary experiment on biomedical verb classification (Korhonen et al., 2006) employed basic syntactic features only: SCFs extracted from corpus data using the system of Briscoe and Carroll (1997) which operates on the output of a domain-independent robust statistical parser (RASP) (Briscoe and Carroll, 2002). Because such deep syntactic features seem ideally suited for challenging biomedical data, we adopted the same basic approach, but we designed and extracted a range of novel feature sets which include additional syntactic and semantic information.

The SCF extraction system assigns each occurrence of a verb in the parsed data as a member of one of the 163 verbal SCFs, builds a lexical entry for each verb (type) and SCF combination, and filters noisy entries out of the lexicon. We do not employ the filter in our work because its primary aim is to filter out SCFs containing adjuncts (as opposed to arguments). Adjuncts have been shown to be beneficial for general language verb classification (Sun et al., 2008; Joanis et al., 2007) and particularly meaningful in biomedical texts (Cohen and Hunter, 2006).

The lexical entries provide various information useful for verb classification, including e.g. the frequency of the entry in the data, the part-of-speech (POS) tags of verb tokens, the argument heads in argument positions, the prepositions in PP frames, and the number of verbal occurrences in active and passive. Making use of this information we designed ten feature sets for experimentation.

The first three feature sets F1-F3 include basic SCF frequency information for each verb:

F1: SCFs and their relative frequencies. The SCFs abstract over lexically governed particles and prepositions.

F2: F1 with two high frequency PP frames parameterized for prepositions: the simple PP and NP-PP frames refined according to the prepositions provided in the lexical entries (e.g. PP_*at*, PP_*on*, PP_*in*).

[2] See section 5 for discussion on previous work.

F3: F2 with 13 additional high frequency PP frames parameterized for prepositions.

Although prepositions are an important part of the syntactic description of lexical classes and therefore F3 should be the most informative feature set, we controlled the number of PP frames parameterized for prepositions to examine the effect of sparse data in automatic classification.

F4-F7 build on the most refined SCF-based feature set F3, supplementing it with information about verb tense (F4-F5) and voice (F6-F7):

F4: The frequencies of POS tags (e.g. VVD for *activated*) calculated over all the SCFs of the verb.

F5: The frequencies of POS tags calculated specific to each SCF of the verb.

F6: The frequency of the active and passive occurrences of the verb (calculated over all the SCFs of the verb).

F7: The frequency of the active and passive occurrences of the verb (calculated specific to each SCF of the verb).

Also F8-F10 build on feature set F3. They supplement it with information about lexical or semantic selectional preferences (SPs) of the verbs in the following slots: subject, direct object, second object, and the NP within the PP complement. The SPs are acquired using argument head data in the ten most frequent SCFs. We use two baseline methods (F8 and F9) which employ raw data and one method based on clustering (F10):

F8: The raw argument head types are considered as SP classes.

F9: Only those raw argument head types which occur with four or more verbs with frequency of ≥ 3 are considered as SP classes.

F10: SPs are acquired by clustering those argument heads which occur with ten or more verbs with frequency of ≥ 3. We used the PC clustering method described below in section 2. The number of clusters K_{np} was set to 10, 20, and 50 to produce SP classes. We call the feature sets corresponding to these different values of K_{np} F10A, F10B and F10C, respectively. Since the clustering algorithms have an element of randomness, clustering was ran 100 times. The output is a result of voting among the outputs of the runs.

F3-F10 are entirely novel feature sets in biomedical verb classification. Variations of some of them have been used in earlier work on general language classification (see section 5 for details).

2.2 Classification

The clustering method which proved the best in the preliminary experiment on biomedical verb classification was Information Bottleneck (IB) (Tishby et al., 1999). We compare this method against a probabilistic method: a cost-based framework for pairwise clustering (PC) (Puzicha et al., 2000).

2.2.1 Information Bottleneck

IB is an information-theoretic method which controls the balance between: (i) the *loss* of information by representing verbs as clusters ($I(Clusters; Verbs)$), which has to be minimal, and (ii) the *relevance* of the output clusters for representing the SCF distribution ($I(Clusters; \text{SCF}s)$) which has to be maximal. The balance between these two quantities ensures optimal compression of data through clusters. The trade-off between the two constraints is realized through minimising the cost function:

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

where β is a parameter that balances the constraints. IB takes three inputs: (i) SCF-verb -based distributions, (ii) the desired number of clusters $\mathcal{K}$, and (iii) the initial value of β. It then looks for the minimal β that decreases $\mathcal{L}_{\text{IB}}$ compared to its value with the initial β, using the given $\mathcal{K}$. IB delivers as output the probabilities $p(K|V)$.

2.2.2 Pairwise Clustering

PC is a method where a cost criterion guides the search for a suitable clustering configuration. This criterion is realized through a cost function $H(S, M)$ where

(i) $S = \{\text{sim}(a, b)\}$, $a, b \in A$: a collection of pairwise similarity values, each of which pertains to a pair of data elements $a, b \in A$.

(ii) $M = (A_1, \ldots, A_k)$: a candidate clustering configuration, specifying assignments of all elements into the disjoint clusters (that is $\cup A_j = A$ and $A_j \cap A_{j'} = \phi$ for every $1 \leq j < j' \leq k$).

1 Have an effect on activity (BIO/29)	9 Report (GEN/30)
1.1 Activate / Inactivate	**9.1 Investigate**
1.1.1 Change activity: *activate, inhibit*	9.1.1 Examine: *evaluate, analyze*
1.1.2 Suppress: *suppress, repres s*	9.1.2 Establish: *test, investigate*
1.1.3 Stimulate: *stimulate*	9.1.3 Confirm: *verify, determine*
1.1.4 Inactivate: *delay, diminish*	**9.2 Suggest**
1.2 Affect	9.2.1 Presentational:
1.2.1 Modulate: *stabilize, modulate*	*hypothesize, conclude*
1.2.2 Regulate: *control, support*	9.2.2 Cognitive:
1.3 Increase / decrease: *increase, decrease*	*consider, believe*
1.4 Modify: *modify, catalyze*	**9.3 Indicate:** *demonstrate, imply*

Table 1: Sample classes from the gold standard

Journal	Years	Words
Genes & Development	2003-5	4.7M
Journal of Biological Chemistry	2004 (Vol.1-9)	5.2M
The Journal of Cell Biology	2003-5	5.6M
Cancer Research	2005	6.5M
Carcinogenesis	2003-5	3.4M
Nature Immunology	2003-5	2.3M
Drug Metabolism and Disposition	2003-5	2.3M
Toxicological Sciences	2003-5	3.1M
Total:		33.1M

Table 2: Data from MEDLINE

The cost function is defined as follows:

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

where n_j is the size of the j^{th} cluster and Avgsim_j is the average similarity between cluster members. We used the Jensen-Shannon divergence (JS) as the similarity measure.

3 Data

3.1 Test Verbs and Gold Standard

We employed in our experiments the same gold standard as earlier employed by Korhonen et al. (2006). This three level gold standard was created by a team of human experts: 4 domain experts and 2 linguists. It includes 192 test verbs (typically frequent verbs in biomedical journal articles) classified into 16, 34 and 50 classes, respectively. The classes created by domain experts are labeled as BIO and those created by linguists as GEN. BIO classes include 116 verbs whose analysis required domain knowledge (e.g. *activate, solubilize, harvest*). GEN classes include 76 general or scientific text verbs (e.g. *demonstrate, hypothesize, appear*). Each class is associated with 1-30 member verbs. Table 1 illustrates two of the gold standard classes with 1-2 example verbs per (sub-)class.

3.2 Test Data

We downloaded the data from the MEDLINE database, from eight journals covering various areas of biomedicine. The first column in table 2 lists each journal, the second shows the years from which the articles were downloaded, and the third indicates the size of the data. We experimented with two test sets: 1) The 15.5M word sub-set shown in the first three rows of the table (this was used for creating the gold standard). 2) All the data: this new larger data was necessary for experiments with new feature sets as the most refined ones do not appear in 1) with sufficient frequency.

SCF	F1	98	39
	F2	247	64
	F3	486	75
F3 + tense	F4	490	79
	F5	920	176
F3 + voice	F6	488	77
	F7	682	153
F3 + SP	F8	150407	2112
	F9	13352	344
	F10A	110280	2091
	F10B	115208	2091
	F10C	114793	2091

Table 3: (i) The total number of features and (ii) the average per verb for all the feature sets

4 Experimental Evaluation

4.1 Processing the Data

The data was first processed using the feature extraction module. Table 3 shows (i) the total number of features in each feature set and (ii) the average per verb in the resulting lexicon. The classification module was then applied. We requested $\mathcal{K} = 2$ to 60 clusters from both clustering methods. We did not want to enforce the actual number of classes but preferred to let the class hierarchy emerge from the clustering results. In order to find the values of $\mathcal{K}$ where the clustering output might correspond to a level in the class hierarchy we used the relevance criterion. For each method (clustering method and feature set combination) we choose as informative $\mathcal{K}$'s the values for which the relevance information $I(Clusters; \text{SCF}s)$) increases more sharply between $\mathcal{K} - 1$ and $\mathcal{K}$ clusters than between $\mathcal{K}$ and $\mathcal{K} + 1$. We then chose for evaluation the outputs corresponding only to informative values of $\mathcal{K}$. The clustering was run 50 times for each method. The output is a result of voting among the outputs of the runs.

4.2 Measures

The clusters were evaluated against the gold standard using four methods. The first measure, the *adjusted pairwise precision*, evaluates clusters in terms of verb pairs:

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

APP is the average proportion of all within-cluster pairs that are correctly co-assigned. Multiplied by a factor that increases with cluster size it compensates for a bias towards small clusters.

The second measure is *modified purity*, 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 $n_{\text{prevalent}}(K)$. Verbs that do not take it are considered as errors. Clusters where $n_{\text{prevalent}}(K) = 1$ are disregarded as not to introduce a bias towards singletons:

$$m\text{PUR} = \frac{\sum\limits_{n_{\text{prevalent}}(k_i) \geq 2} n_{\text{prevalent}}(k_i)}{\text{number of verbs}}$$

The third measure is the *weighted class accuracy*, the proportion of members of dominant clusters DOM-CLUST_i within all classes c_i.

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

mPUR can be seen to measure the precision of clusters and ACC the recall. We define an F measure as the harmonic mean of mPUR and ACC:

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

The experiments were run 50 times on each input to get the distribution of performance due to the randomness in the initial clustering. We calculated the average performance and standard deviation from the results of these runs.

4.3 Results for Test Set 1

We first compared IB and PC on the smaller test set 1 using feature set F2. We chose for evaluation the outputs corresponding to the most informative values of $\mathcal{K}$: 20, 33, 53 for IB, and 19, 26, 51 for PC. In the results included in table 4 IB shows slightly better performance than PC, but the difference is not significant for K=34 and 50. We decided to use PC for larger experiments because it has two advantages over IB: 1) It can cluster the large test set 2 with $\mathcal{K} = 10 - 60$ in minutes, while IB requires a day for this. 2) It can deal with (and combine) different feature sets, while IB runs into numerical problems. Due to its speed and flexibility PC is thus more suitable for larger-scale experiments involving comparison of complex feature sets.

4.4 Results for Test Set 2

Tables 5 and 6 include the PC results on the larger test set 2. Table 5 shows the results for each individual feature set (indicated in the second column). It shows also the standard deviations (σ_{avg}) of the four performance measures averaged across all the runs. These are very similar for 16, 34, and 50 classes and hence only included in one of the columns. In addition, σ_{diff} is indicated. This is $\sqrt{2} \cdot \sigma_{\text{avg}}$ and used for calculating the significance of the performance differences. In the following discussion we consider a difference of more than $2\sigma_{\text{diff}}$ ($p > 97.7\%$) as significant.

The first feature sets F1-F3 include basic SCF (frequency) information for each verb, F2-F3 refined with prepositions. F2 shows clearly better results than F1 (over 10 F-measure) at all the levels of gold standard. This demonstrates the usefulness of prepositions for the task. When moving to F3 the performance decreases for 34 and 50 classes, while improving for 16 classes, but these differences are not statistically significant.

Feature sets F4-F10 build on F3. F4-F5 include information about verb tense. This information proves quite useful for verb classification, particularly when specific to individual SCFs. When compared against the baseline featureset F3, F5 is clearly better - particularly at 50 classes where the difference is 3.9 in F-measure ($2\sigma_{\text{diff}}$). Verb voice information is not equally helpful: F6-F7 are not better than F3. In some comparisons they are worse, e.g. F7 vs. F3 at 16 classes.

F8-F10 supplement F3 with information about SPs. Surprisingly, these lexical and semantic features prove the most useful for our task. At the level of 34 and 50 classes, the best SP features are even better than the best tense features (the difference is statistically significant), and they yield notable improvement over the baseline features (e.g. 6.8 difference in F-measure between F9 and F3). The performance is not equally good at 16 classes. This makes perfect sense because class members are unlikely to have similar SPs at such a coarse level of semantic classification.

When comparing the five sets of SPs features against each other, F9 and F10C produce the best results at 34 and 50 classes. F9 uses raw (filtered) argument head data for SP acquisition while F10C uses clustering. It is interesting that the difference between these two very different methods is not statistically significant. Whether one employs

	16 Classes				34 Classes				50 Classes			
	APP	mPUR	ACC	F	APP	mPUR	ACC	F	APP	mPUR	ACC	F
IB	74	77	66	71	69	75	81	77	54	72	79	75
PC	71	78	58	67	64	71	81	75	63	71	73	72
±	1.1	1.0	1.0	0.8	1.8	1.6	1.3	1.4	2.1	1.5	1.6	1.1

Table 4: Performance on test set 1

		16 Classes				34 Classes				50 Classes			
		APP	mPUR	ACC	F	APP	mPUR	ACC	F	APP	mPUR	ACC	F
SCF	F1	62.7	68.2	54.6	60.6	50.4	58.4	53.4	55.8	41.5	50.3	55.7	52.9
	F2	68.7	76.4	66.4	71.1	61.9	65.5	65.8	65.6	53.9	61.2	65.4	63.2
	F3	69.3	77.7	67.6	72.3	61.6	66.0	64.0	65.0	53.7	60.2	65.9	62.9
F3 + tense	F4	70.1	77.5	65.5	71.0	62.0	70.3	69.4	69.8	53.3	60.6	68.0	64.1
	F5	68.5	75.4	71.7	73.5	61.9	67.8	68.2	68.0	58.2	62.7	71.7	66.8
F3 + voice	F6	70.6	78.1	64.0	70.4	61.2	66.0	65.8	65.9	54.3	59.6	70.1	64.4
	F7	74.0	79.5	59.7	68.2	62.6	65.4	65.1	65.2	55.1	60.9	69.2	64.7
F3 + SP	F8	77.1	78.2	61.6	68.9	69.6	69.3	71.2	70.2	61.3	62.7	71.1	66.6
	F9	72.4	77.1	64.0	69.9	72.2	72.0	71.6	71.8	62.3	65.6	72.4	68.8
	F10A	75.6	80.0	63.2	70.6	66.1	69.2	70.6	69.9	59.4	63.5	69.0	66.2
	F10B	68.8	77.1	69.2	72.9	65.3	67.2	69.8	68.5	59.9	61.9	70.5	65.9
	F10C	74.1	78.9	65.7	71.7	68.8	71.7	69.7	70.7	59.8	63.4	71.1	67.0
	σ_{avg}	2.2	1.5	1.8	1.4								
	σ_{diff}	3.1	2.1	2.5	2.0								

Table 5: Performance on test set 2: PC clustering results for individual feature sets at the three levels of gold standard. σ_{avg} and σ_{diff} were calculated across all the three classification levels.

16 CL.	F5+F9	F4+ F10C	F5	F5+ F8
APP	72.3	68.2	68.5	72.2
mPUR	76.4	77.0	75.4	76.5
ACC	73.6	70.9	71.7	69.9
F	75.0	73.8	73.5	73.0
34 CL.	F5+ F9	F5+ F8	F9	F4+ F10A
APP	68.7	71.0	72.2	62.9
mPUR	70.1	71.0	72.0	68.4
ACC	74.8	73.4	71.6	75.0
F	72.4	72.2	71.8	71.5
50 CL.	F9	F5+ F9	F5+ F8	F4+ F9
APP	62.3	59.8	62.8	59.7
mPUR	65.6	63.8	64.1	63.1
ACC	72.4	72.7	71.0	71.8
F	68.8	68.0	67.4	67.1

Table 6: Results for the top four feature set combinations. All the feature sets build on F3.

fine grained clusters (F10C) or coarse-grained ones (F10A) as SPs does not make much difference.

We next combined various feature sets. Table 6 shows the performance for the top four combinations. Comparing these results against the ones in Table 5, (see the σ_{diff} values in Table 5) we can see that combining feature sets does not result in better performance[3]. The only exception is the difference in APP and mPUR between F9 and F4 + F10A at N=34. However, these results show similar tendencies as the earlier ones: at 16 classes the most useful features are based on verb tense, while at 34 and 50 classes they are based on SPs.

5 Discussion

The results presented in the previous section are in interesting contrast with those reported in earlier work. In previous work on general language verb classification, syntactic features (slots or SCFs) have proved generally the most helpful features, e.g. (Schulte im Walde, 2006; Joanis et al., 2007). The preliminary experiment on biomedical verb classification (Korhonen et al., 2006) experimented only with them. In our experiments, SCFs proved useful baseline features. When we refined them further, we faced sparse data problems: considerable improvement was obtained when moving from F1 to F2, but not when moving to F3. Although many verb classes are sensitive to preposition types, many of the types are low in frequency. Future work could address this problem by employing smoothing techniques, or backing off to preposition classes.

Joanis et al. (2007) experimented with tense and voice -based features in general English verb classification. They offered no significant improvement over basic syntactic features. Also in our experiments, we obtained little improvement with voice features. This could be due to the

[3] Recall that all F4-F10 are actually already ’combined’ with F3 - we do not refer to this combination here.

un-distinctiveness of passive in biomedical texts where it is used typically with high frequency. However, tense-based features clearly improved the baseline performance in our experiments. This could be partly because we 'parameterize' POS information for SCFs, and partly because semantically similar verbs in biomedical language tend to behave similarly also in terms of tense (Friedman et al., 2002).

Joanis (2002) and Schulte im Walde (2006) used SP-based features in general English and German verb classifications, respectively. The former acquired them from WordNet (Miller, 1990) and the latter from GermaNet (Kunze, 2000). Joanis (2002) obtained no improvement over syntactic features while Schulte im Walde (2006) obtained, but the improvement was not significant. In our experiments, SP features gave the best results and the clearest improvement over the baseline features at the finer-grained levels of classification where class members are indeed likely to be the most uniform in terms of their SPs.

We obtained this improvement despite using a fully unsupervised approach to SP acquisition. We did not exploit lexical resources like Joanis (2002) and Schulte im Walde (2006) because it would have required combining general resources (e.g. WordNet) with domain specific ones (e.g. UMLS). We opted for a simpler approach in this initial work – using raw argument heads and clustering – and obtained surprisingly good results. In our experiments filtering of raw argument heads and clustering with N=50 produced equivalent results, suggesting that relatively fine-grained clusters are optimal. Future work will require qualitative analysis of noun clusters and comparison of these against classes in lexical resources to determine an optimal method for SP acquisition.

Does the fact that we obtain good results with features which have not proved helpful in general language classification indicate a need for domain-specific feature engineering? We do not believe so. The feature sets we experimented with are theoretically well-motivated and should, in principle, also aid general language verb classification. We believe they proved helpful in our experiments because being domain-specific, biomedical language is conventionalised and therefore less varied in terms of verb sense and usage than general language. For example, verbs have stronger SPs for their argument heads when many of their corpus occurrences are of similar sense. This renders SP-based features more useful for classification.

Due to differences in the data, methods, and experimental setup, direct comparison of our performance figures with previously published ones is difficult. The closest comparison point with general language is (Korhonen et al., 2003) which reported 59% mPUR using IB to assign 110 polysemous English verbs into 34 classes. Our best results are substantially better (72-80% mPUR). It is encouraging that we obtained such good results despite focusing on a linguistically challenging domain.

In addition to the points mentioned earlier, our future plans include seeding automatic classification with more sophisticated information acquired automatically from domain-specific texts (e.g. using named entity recognition and anaphoric linking (Vlachos et al., 2006)). We will also explore semi-automatic ML technology and active learning in aiding the classification. Finally, we plan to conduct a bigger experiment with a larger number of verbs, make the resulting classification publicly available, and demonstrate its usefulness for practical BIO-NLP application tasks.

6 Conclusion

We reported large-scale experiments to investigate the optimal characteristics of features required for biomedical verb classification. A range of feature sets and associated extraction methods were introduced for this work, along with a robust clustering method capable of dealing with large data and complex feature sets. A number of experiments were reported. The best performing feature sets proved to be the ones which include information about SCFs supplemented with information about verb tense and SPs in particular. The latter were acquired automatically from corpus data using an unsupervised method. Similar feature sets have not proved equally useful in earlier work in general language verb classification. We discussed reasons for this and highlighted several areas for future work.

Acknowledgement

Work on this paper was funded by the Royal Society, EPSRC ('ACLEX' project, GR/T19919/01) and MRC ('CRAB' project, G0601766), UK.

References

Ananiadou, S., B. D. Kell, and J. Tsujii. 2006. Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12):571–579.

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

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

Cohen, K. B. and L. Hunter. 2006. A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics*, 7(3).

Dang, H. T. 2004. *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation*. Ph.D. thesis, CIS, University of Pennsylvania.

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

Friedman, C., P. Kra, and A. Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Joanis, E., S. Stevenson, and D. James. 2007. A general feature space for automatic verb classification. *Natural Language Engineering*.

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

Korhonen, A., Y. Krymolowski, and N. Collier. 2006. Automatic classification of verbs in biomedical texts. In *ACL-COLING*, Sydney, Australia.

Kunze, C. 2000. Extension and use of germanet, a lexical-semantic database. In *2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

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

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

Prescher, D., S. Riezler, and M. Rooth. 2000. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *18th International Conference on Computational Linguistics*, pages 649–655, Saarbrücken, Germany.

Puzicha, J., T. Hofmann, and J. M. Buhmann. 2000. A theory of proximity-based clustering: structure detection by optimization. *Pattern Recognition*, 33(4):617–634.

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

Shi, L. and R. Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.

Sun, L., A. Korhonen, and Y. Krymolowski. 2008. Verb class discovery from rich syntactic data. In *9th International Conference on Intelligent Text Processing and Computational Linguistics*, Haifa, Israel.

Swier, R. and S. Stevenson. 2004. Unsupervised semantic role labelling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 95–102, Barcelona, Spain, August.

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

Vlachos, A., C. Gasperin, I. Lewin, and E. J. Briscoe. 2006. Bootstrapping the recognition and anaphoric linking of named entitites in drosophila articles. In *Pacific Symposium in Biocomputing*, Maui, Hawaii.