# Automatic Classification of Verbs in Biomedical Texts

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

**Yuval Krymolowski**
Dept. of Computer Science
Technion
Haifa 32000
Israel
yuvalkr@cs.technion.ac.il

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

## Abstract

Lexical classes, when tailored to the application and domain in question, can provide an effective means to deal with a number of natural language processing (NLP) tasks. While manual construction of such classes is difficult, recent research shows that it is possible to automatically induce verb classes from cross-domain corpora with promising accuracy. We report a novel experiment where similar technology is applied to the important, challenging domain of biomedicine. We show that the resulting classification, acquired from a corpus of biomedical journal articles, is highly accurate and strongly domain-specific. It can be used to aid BIO-NLP directly or as useful material for investigating the syntax and semantics of verbs in biomedical texts.

## 1 Introduction

Lexical classes which capture the close relation between the syntax and semantics of verbs have attracted considerable interest in NLP (Jackendoff, 1990; Levin, 1993; Dorr, 1997; Prescher et al., 2000). Such classes are useful for their ability to capture generalizations about a range of linguistic properties. For example, verbs which share the meaning of 'manner of motion' (such as *travel*, *run*, *walk*), behave similarly also in terms of subcategorization (*I traveled/ran/walked, I traveled/ran/walked to London, I traveled/ran/walked five miles*). Although the correspondence between the syntax and semantics of words is not perfect and the classes do not provide means for full semantic inferencing, their predictive power is nevertheless considerable.

NLP systems can benefit from lexical classes in many ways. Such classes define the mapping from surface realization of arguments to predicate-argument structure, and are therefore an important component of any system which needs the latter. As the classes can capture higher level abstractions they can be used as a means to abstract away from individual words when required. They are also helpful in many operational contexts where lexical information must be acquired from small application-specific corpora. Their predictive power can help compensate for lack of data fully exemplifying the behavior of relevant words.

Lexical verb classes have been used to support various (multilingual) tasks, such as computational lexicography, language generation, machine translation, word sense disambiguation, semantic role labeling, and subcategorization acquisition (Dorr, 1997; Prescher et al., 2000; Korhonen, 2002). However, large-scale exploitation of the classes in real-world or domain-sensitive tasks has not been possible because the existing classifications, e.g. (Levin, 1993), are incomprehensive and unsuitable for specific domains.

While manual classification of large numbers of words has proved difficult and time-consuming, recent research shows that it is possible to automatically induce lexical classes from corpus data with promising accuracy (Merlo and Stevenson, 2001; Brew and Schulte im Walde, 2002; Korhonen et al., 2003). A number of ML methods have been applied to classify words using features pertaining to mainly syntactic structure (e.g. statistical distributions of subcategorization frames (SCFs) or general patterns of syntactic behaviour, e.g. transitivity, passivisability) which have been extracted from corpora using e.g. part-of-speech tagging or robust statistical parsing techniques.



*Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 345–352,
Sydney, July 2006. 

This research has been encouraging but it has so far concentrated on general language. Domain-specific lexical classification remains unexplored, although it is arguably important: existing classifications are unsuitable for domain-specific applications and these often challenging applications might benefit from improved performance by utilizing lexical classes the most.

In this paper, we extend an existing approach to lexical classification (Korhonen et al., 2003) and apply it (without any domain specific tuning) to the domain of biomedicine. We focus on biomedicine for several reasons: (i) NLP is critically needed to assist the processing, mining and extraction of knowledge from the rapidly growing literature in this area, (ii) the domain lexical resources (e.g. UMLS metathesaurus and lexicon[1]) do not provide sufficient information about verbs and (iii) being linguistically challenging, the domain provides a good test case for examining the potential of automatic classification.

We report an experiment where a classification is induced for 192 relatively frequent verbs from a corpus of 2230 biomedical journal articles. The results, evaluated with domain experts, show that the approach is capable of acquiring classes with accuracy higher than that reported in previous work on general language. We discuss reasons for this and show that the resulting classes differ substantially from those in extant lexical resources. They constitute the first syntactic-semantic verb classification for the biomedical domain and could be readily applied to support BIO-NLP.

We discuss the domain-specific issues related to our task in section 2. The approach to automatic classification is presented in section 3. Details of the experimental evaluation are supplied in section 4. Section 5 provides discussion and section 6 concludes with directions for future work.

## 2 The Biomedical Domain and Our Task

Recent years have seen a massive growth in the scientific literature in the domain of biomedicine. For example, the MEDLINE database[2] which currently contains around 16M references to journal articles, expands with 0.5M new references each year. 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 NLP tools which can be used to automatically locate, organize and manage facts related to published experimental results.

In recent years, 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 (Hirschman et al., 2002). Other tasks, such as the extraction of factual information, remain a bigger challenge. This is partly due to the challenging nature of biomedical texts. They are complex both in terms of syntax and semantics, containing complex nominals, modal subordination, anaphoric links, etc.

Researchers have recently began to use deeper NLP techniques (e.g. statistical parsing) in the domain because they are not challenged by the complex structures to the same extent than shallow techniques (e.g. regular expression patterns) are (Lease and Charniak, 2005). However, deeper techniques require richer domain-specific lexical information for optimal performance than is provided by existing lexicons (e.g. UMLS). This is particularly important for verbs, which are central to the structure and meaning of sentences.

Where the lexical information is absent, lexical classes can compensate for it or aid in obtaining it in the ways described in section 1. Consider e.g. the INDICATE and ACTIVATE verb classes in Figure 1. They capture the fact that their members are similar in terms of syntax and semantics: they have similar SCFs and selectional preferences, and they can be used to make similar statements which describe similar events. Such information can be used to build a richer lexicon capable of supporting key tasks such as parsing, predicate-argument identification, event extraction and the identification of biomedical (e.g. interaction) relations.

While an abundance of work has been conducted on semantic classification of biomedical terms and nouns, less work has been done on the (manual or automatic) semantic classification of verbs in the biomedical domain (Friedman et al., 2002; Hatzivassiloglou and Weng, 2002; Spasic et al., 2005). No previous work exists in this domain on the type of *lexical* (i.e. syntactic-semantic) verb classification this paper focuses on.

To get an initial idea about the differences between our target classification and a general language classification, we examined the extent to which individual verbs and their frequencies differ in biomedical and general language texts. We

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

[2] http://www.ncbi.nlm.nih.gov/PubMed/

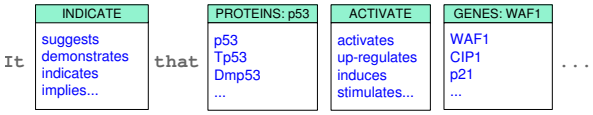


Figure 1: Sample lexical classes

| BIO | BNC |
|---|---|
| *show* | *do* |
| *suggest* | *say* |
| *use* | *make* |
| *indicate* | *go* |
| *contain* | *see* |
| *describe* | *take* |
| *express* | *get* |
| *bind* | *know* |
| *require* | *come* |
| *observe* | *give* |
| *find* | *think* |
| *determine* | *use* |
| *demonstrate* | *find* |
| *perform* | *look* |
| *induce* | *want* |

Table 1: The 15 most frequent verbs in the biomedical data and in the BNC

created a corpus of 2230 biomedical journal articles (see section 4.1 for details) and compared the distribution of verbs in this corpus with that in the British National Corpus (BNC) (Leech, 1992). We calculated the Spearman rank correlation between the 1165 verbs which occurred in both corpora. The result was only a weak correlation: $0.37 \pm 0.03$. When the scope was restricted to the 100 most frequent verbs in the biomedical data, the correlation was $0.12 \pm 0.10$ which is only $1.2\sigma$ away from zero. The dissimilarity between the distributions is further indicated by the Kullback-Leibler distance of 0.97. Table 1 illustrates some of these big differences by showing the list of 15 most frequent verbs in the two corpora.

## 3 Approach

We extended the system of Korhonen et al. (2003) with additional clustering techniques (introduced in sections 3.2.2 and 3.2.4) and used it to obtain the classification for the biomedical domain. The system (i) extracts features from corpus data and (ii) clusters them using five different methods. These steps are described in the following two sections, respectively.

### 3.1 Feature Extraction

We employ as features distributions of SCFs specific to given verbs. We extract them from corpus data using the comprehensive subcategorization acquisition system of Briscoe and Carroll (1997) (Korhonen, 2002). The system incorporates RASP, a domain-independent robust statistical parser (Briscoe and Carroll, 2002), which tags, lemmatizes and parses data yielding complete though shallow parses and a SCF classifier which incorporates an extensive inventory of 163 verbal SCFs[3]. The SCFs abstract over specific lexically-governed particles and prepositions and specific predicate selectional preferences. In our work, we parameterized two high frequency SCFs for prepositions (PP and NP + PP SCFs). No filtering of potentially noisy SCFs was done to provide clustering with as much information as possible.

### 3.2 Classification

The SCF frequency distributions constitute the input data to automatic classification. We experiment with five clustering methods: the simple hard nearest neighbours method and four probabilistic methods – two variants of Probabilistic Latent Semantic Analysis and two information theoretic methods (the Information Bottleneck and the Information Distortion).

#### 3.2.1 Nearest Neighbours

The first method collects the nearest neighbours (NN) of each verb. It (i) calculates the Jensen-Shannon divergence (JS) between the SCF distributions of each pair of verbs, (ii) connects each verb with the most similar other verb, and finally (iii) finds all the connected components. The NN method is very simple. It outputs only one clustering configuration and therefore does not allow examining different cluster granularities.

#### 3.2.2 Probabilistic Latent Semantic Analysis

The Probabilistic Latent Semantic Analysis (PLSA, Hoffman (2001)) assumes a generative model for the data, defined by selecting (i) a verb $verb_i$, (ii) a semantic class $class_k$ from the distribution $p(Classes \,|\, verb_i)$, and (iii) a SCF $scf_j$ from the distribution $p(\mathrm{SCFs} \,|\, class_k)$. PLSA uses Expectation Maximization (EM) to find the distribution $\tilde{p}(\mathrm{SCFs} \,|\, Clusters, Verbs)$ which maximises the likelihood of the observed counts. It does this by minimising the cost function

$$\mathcal{F} = -\beta \log \mathrm{Likelihood}(\tilde{p} \,|\, \mathrm{data}) + H(\tilde{p}) \,.$$

[3] See http://www.cl.cam.ac.uk/users/alk23/subcat/subcat.html for further detail.

For $\beta = 1$ minimising $\mathcal{F}$ is equivalent to the standard EM procedure while for $\beta < 1$ the distribution $\tilde{p}$ tends to be more evenly spread. We use $\beta = 1$ (PLSA/EM) and $\beta = 0.75$ (PLSA$_{\beta=0.75}$). We currently "harden" the output and assign each verb to the most probable cluster only[4].

#### 3.2.3 Information Bottleneck

The Information Bottleneck (Tishby et al., 1999) (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:

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

where $\beta$ is a parameter that balances the constraints. IB takes three inputs: (i) SCF-verb distributions, (ii) the desired number of clusters $\mathcal{K}$, and (iii) the initial value of $\beta$. It then looks for the minimal $\beta$ that decreases $\mathcal{L}_{\text{IB}}$ compared to its value with the initial $\beta$, using the given $\mathcal{K}$. IB delivers as output the probabilities $p(K|V)$. It gives an indication for the most informative number of output configurations: the ones for which the relevance information increases more sharply between $\mathcal{K}-1$ and $\mathcal{K}$ clusters than between $\mathcal{K}$ and $\mathcal{K}+1$.

#### 3.2.4 Information Distortion

The Information Distortion method (Dimitrov and Miller, 2001) (ID) is otherwise similar to IB but $\mathcal{L}_{\text{ID}}$ differs from $\mathcal{L}_{\text{IB}}$ by an additional term that adds a bias towards clusters of similar size:

$$\begin{aligned}\mathcal{L}_{\text{ID}} &= -H(Clusters\,|\,Verbs) \\ &\quad - \beta I(Clusters; \text{SCF}s) \\ &= \mathcal{L}_{\text{IB}} - H(Clusters)\,.\end{aligned}$$

ID yields more evenly divided clusters than IB.

## 4 Experimental Evaluation

### 4.1 Data

We downloaded the data for our experiment from the MEDLINE database, from three of the 10 leading journals in biomedicine: 1) *Genes & Development* (molecular biology, molecular genetics), 2) *Journal of Biological Chemistry* (biochemistry and molecular biology) and 3) *Journal of Cell Biology* (cellular structure and function). 2230 full-text articles from years 2003-2004 were used. The data included 11.5M words and 323,307 sentences in total. 192 medium to high frequency verbs (with the minimum of 300 occurrences in the data) were selected for experimentation[5]. This test set was big enough to produce a useful classification but small enough to enable thorough evaluation in this first attempt to classify verbs in the biomedical domain.

### 4.2 Processing the Data

The data was first processed using the feature extraction module. 233 (preposition-specific) SCF types appeared in the resulting lexicon, 36 per verb on average.[6] The classification module was then applied. NN produced $\mathcal{K}_{\text{nn}} = 42$ clusters. From the other methods we requested $\mathcal{K} = 2$ to 60 clusters. We chose for evaluation the outputs corresponding to the most informative values of $\mathcal{K}$: 20, 33, 53 for IB, and 17, 33, 53 for ID.

### 4.3 Gold Standard

Because no target lexical classification was available for the biomedical domain, human experts (4 domain experts and 2 linguists) were used to create the gold standard. They were asked to examine whether the test verbs similar in terms of their syntactic properties (i.e. verbs with similar SCF distributions) are similar also in terms of semantics (i.e. they share a common meaning). Where this was the case, a verb class was identified and named.

The domain experts examined the 116 verbs whose analysis required domain knowledge (e.g. *activate, solubilize, harvest*), while the linguists analysed the remaining 76 general or scientific text verbs (e.g. *demonstrate, hypothesize, appear*). The linguists used Levin (1993) classes as gold standard classes whenever possible and created novel ones when needed. The domain experts used two purely semantic classifications of biomedical verbs (Friedman et al., 2002; Spasic et al., 2005)[7] as a starting point where this was pos-

[4] The same approach was used with the information theoretic methods. It made sense in this initial work on biomedical classification. In the future we could use soft clustering a means to investigate polysemy.

[5] 230 verbs were employed initially but 38 were dropped later so that each (coarse-grained) class would have the minimum of 2 members in the gold standard.

[6] This number is high because no filtering of potentially noisy SCFs was done.

[7] See http://www.cbr-masterclass.org.

| | |
|---|---|
| 1 Have an effect on activity (BIO/29) | 8 Physical Relation Between Molecules (BIO/20) |
| **1.1 Activate / Inactivate** | **8.1 Binding:** *bind, attach* |
| 1.1.1 Change activity: *activate, inhibit* | **8.2 Translocate and Segregate** |
| 1.1.2 Suppress: *suppress, repress* | 8.2.1 Translocate: *shift, switch* |
| 1.1.3 Stimulate: *stimulate* | 8.2.2 Segregate: *segregate, export* |
| 1.1.4 Inactivate: *delay, diminish* | **8.3 Transmit** |
| **1.2 Affect** | 8.3.1 Transport: *deliver, transmit* |
| 1.2.1 Modulate: *stabilize, modulate* | 8.3.2 Link: *connect, map* |
| 1.2.2 Regulate: *control, support* | 9 Report (GEN/30) |
| **1.3 Increase / decrease:** *increase, decrease* | **9.1 Investigate** |
| **1.4 Modify:** *modify, catalyze* | 9.1.1 Examine: *evaluate, analyze* |
| 2 Biochemical events (BIO/12) | 9.1.2 Establish: *test, investigate* |
| **2.1 Express:** *express, overexpress* | 9.1.3 Confirm: *verify, determine* |
| **2.2 Modification** | **9.2 Suggest** |
| 2.2.1 Biochemical modification: *dephosphorylate, phosphorylate* | 9.2.1 Presentational: *hypothesize, conclude* |
| 2.2.2 Cleave: *cleave* | 9.2.2 Cognitive: *consider, believe* |
| **2.3 Interact:** *react, interfere* | **9.3 Indicate:** *demonstrate, imply* |
| 3 Removal (BIO/6) | 10 Perform (GEN/10) |
| **3.1 Omit:** *displace, deplete* | **10.1 Quantify** |
| **3.2 Subtract:** *draw, dissect* | 10.1.1 Quantitate: *quantify, measure* |
| 4 Experimental Procedures (BIO/30) | 10.1.2 Calculate: *calculate, record* |
| **4.1 Prepare** | 10.1.3 Conduct: *perform, conduct* |
| 4.1.1 Wash: *wash, rinse* | **10.2 Score:** *score, count* |
| 4.1.2 Mix: *mix* | 11 Release (BIO/4): *detach, dissociate* |
| 4.1.3 Label: *stain, immunoblot* | 12 Use (GEN/4): *utilize, employ* |
| 4.1.4 Incubate: *preincubate, incubate* | 13 Include (GEN/11) |
| 4.1.5 Elute: *elute* | **13.1 Encompass:** *encompass, span* |
| **4.2 Precipitate:** *coprecipitate coimmunoprecipitate* | **13.2 Include:** *contain, carry* |
| **4.3 Solubilize:** *solubilize,lyse* | 14 Call (GEN/3): *name, designate* |
| **4.4 Dissolve:** *homogenize, dissolve* | 15 Move (GEN/12) |
| **4.5 Place:** *load, mount* | **15.1 Proceed:** *progress, proceed* |
| 5 Process (BIO/5): *linearize, overlap* | **15.2 Emerge:** *arise, emerge* |
| 6 Transfect (BIO/4): *inject, microinject* | 16 Appear (GEN/6): *appear, occur* |
| 7 Collect (BIO/6) | |
| **7.1 Collect:** *harvest, select* | |
| **7.2 Process:** *centrifuge, recover* | |

Table 2: The gold standard classification with a few example verbs per class

sible (i.e. where they included our test verbs and also captured their relevant senses)[8].

The experts created a 3-level gold standard which includes both broad and finer-grained classes. Only those classes / memberships were included which all the experts (in the two teams) agreed on.[9] The resulting gold standard including 16, 34 and 50 classes is illustrated in table 2 with 1-2 example verbs per class. The table indicates which classes were created by domain experts (BIO) and which by linguists (GEN). Each class was associated with 1-30 member verbs[10]. The total number of verbs is indicated in the table (e.g. 10 for PERFORM class).

### 4.4 Measures

The clusters were evaluated against the gold standard using measures which are applicable to all the classification methods and which deliver a numerical value easy to interpret.

The first measure, the *adjusted pairwise precision*, evaluates clusters in terms of verb pairs:

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

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:

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

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

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

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

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

The statistical significance of the results is measured by randomisation tests where verbs are swapped between the clusters and the resulting clusters are evaluated. The swapping is repeated 100 times for each output and the average $av_{\text{swaps}}$ and the standard deviation $\sigma_{\text{swaps}}$ is measured. The significance is the scaled difference $signif = (result - av_{\text{swaps}})/\sigma_{\text{swaps}}$ .

### 4.5 Results from Quantitative Evaluation

Table 3 shows the performance of the five clustering methods for $\mathcal{K} = 42$ clusters (as produced by the NN method) at the 3 levels of gold standard classification. Although the two PLSA variants (particularly PLSA$_{\beta=0.75}$) produce a fairly accurate coarse grained classification, they perform worse than all the other methods at the finer-grained levels of gold standard, particularly according to the global measures. Being based on

[8] Purely semantic classes tend to be finer-grained than lexical classes and not necessarily syntactic in nature. Only these two classifications were found to be similar enough to our target classification to provide a useful starting point. Section 5 includes a summary of the similarities/differences between our gold standard and these other classifications.

[9] Experts were allowed to discuss the problematic cases to obtain maximal accuracy - hence no inter-annotator agreement is reported.

[10] The minimum of 2 member verbs were required at the coarser-grained levels of 16 and 34 classes.

| | 16 Classes | | | | 34 Classes | | | | 50 Classes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | APP | $m$PUR | ACC | $F$ | APP | $m$PUR | ACC | $F$ | APP | $m$PUR | ACC | $F$ |
| NN | 81 | 86 | 39 | 53 | 64 | 74 | 62 | 67 | 54 | 67 | 73 | 69 |
| IB | 74 | 88 | 47 | 61 | 61 | 76 | 74 | 75 | 55 | 69 | 87 | 76 |
| ID | 79 | 89 | 37 | 52 | 63 | 78 | 65 | 70 | 53 | 70 | 77 | 73 |
| PLSA/EM | 55 | 72 | 49 | 58 | 43 | 53 | 61 | 57 | 35 | 47 | 66 | 55 |
| PLSA$_{\beta=0.75}$ | 65 | 71 | 68 | 70 | 53 | 48 | 76 | 58 | 41 | 34 | 77 | 47 |

Table 3: The performance of the NN, PLSA, IB and ID methods with $\mathcal{K}_{\mathrm{nn}} = 42$ clusters

| | | 16 Classes | | | | 34 Classes | | | | 50 Classes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{K}$ | | APP | $m$PUR | ACC | $F$ | APP | $m$PUR | ACC | $F$ | APP | $m$PUR | ACC | $F$ |
| 20 | IB | **74** | **77** | **66** | **71** | 60 | 56 | 86 | 67 | 54 | 48 | 93 | 63 |
| 17 | ID | **67** | **76** | **60** | **67** | 43 | 56 | 81 | 66 | 34 | 46 | 91 | 61 |
| 33 | IB | 78 | 87 | 52 | 65 | **69** | **75** | **81** | **77** | 61 | 67 | 93 | 77 |
| | ID | 81 | 88 | 43 | 57 | **65** | **75** | **70** | **72** | 54 | 67 | 82 | 73 |
| 53 | IB | 71 | 87 | 41 | 55 | 61 | 78 | 66 | 71 | **54** | **72** | **79** | **75** |
| | ID | 79 | 89 | 33 | 48 | 66 | 79 | 55 | 64 | **53** | **72** | **68** | **69** |

Table 4: The performance of IB and ID for the 3 levels of class hierarchy for informative values of $\mathcal{K}$

pairwise similarities, NN shows mostly better performance than IB and ID on the pairwise measure APP but the global measures are better for IB and ID. The differences are smaller in $m$PUR (yet significant: $2\sigma$ between NN and IB and $3\sigma$ between NN and ID) but more notable in ACC (which is e.g. $8 - 12\%$ better for IB than for NN). Also the $F$ results suggest that the two information theoretic methods are better overall than the simple NN method.

IB and ID also have the advantage (over NN) that they can be used to produce a hierarchical verb classification. Table 4 shows the results for IB and ID for the informative values of $\mathcal{K}$. The bold font indicates the results when the match between the values of $\mathcal{K}$ and the number of classes at the particular level of the gold standard is the closest.

IB is clearly better than ID at all levels of gold standard. It yields its best results at the medium level (34 classes) with $\mathcal{K} = 33$: $F = 77$ and APP $= 69$ (the results for ID are $F = 72$ and APP $= 65$). At the most fine-grained level (50 classes), IB is equally good according to $F$ with $\mathcal{K} = 33$, but APP is 8% lower. Although ID is occasionally better than IB according to APP and $m$PUR (see e.g. the results for 16 classes with $\mathcal{K} = 53$) this never happens in the case where the correspondence between the number of gold standard classes and the values of $\mathcal{K}$ is the closest. In other words, the informative values of $\mathcal{K}$ prove really informative for IB. The lower performance of ID seems to be due to its tendency to create evenly sized clusters.

All the methods perform significantly better than our random baseline. The significance of the results with respect to two swaps was at the $2\sigma$ level, corresponding to a 97% confidence that the results are above random.

### 4.6 Qualitative Evaluation

We performed further, qualitative analysis of clusters produced by the best performing method IB. Consider the following clusters:

**A:** *inject, transfect, microinfect, contransfect* (6)

**B:** *harvest, select, collect* (7.1)
*centrifuge, process, recover* (7.2)

**C:** *wash, rinse* (4.1.1)
*immunoblot* (4.1.3)
*overlap* (5)

**D:** *activate* (1.1.1)

When looking at coarse-grained outputs, interestingly, $\mathcal{K}$ as low as 8 learned the broad distinction between biomedical and general language verbs (the two verb types appeared only rarely in the same clusters) and produced large semantically meaningful groups of classes (e.g. the coarse-grained classes EXPERIMENTAL PROCEDURES, TRANSFECT and COLLECT were mapped together). $\mathcal{K} = 12$ was sufficient to identify several classes with very particular syntax One of them was TRANSFECT (see **A** above) whose members were distinguished easily because of their typical SCFs (e.g. *inject /transfect/microinfect/contransfect* X *with/into* Y).

On the other hand, even $\mathcal{K} = 53$ could not identify classes with very similar (yet un-identical) syntax. These included many semantically similar sub-classes (e.g. the two sub-classes of COLLECT

shown in **B** whose members take similar NP and PP SCFs). However, also a few semantically different verbs clustered wrongly because of this reason, such as the ones exemplified in **C**. In **C**, *immunoblot* (from the LABEL class) is still somewhat related to *wash* and *rinse* (the WASH class) because they all belong to the larger EXPERIMENTAL PROCEDURES class, but *overlap* (from the PROCESS class) shows up in the cluster merely because of syntactic idiosyncracy.

While parser errors caused by the challenging biomedical texts were visible in some SCFs (e.g. looking at a sample of SCFs, some adjunct instances were listed in the argument slots of the frames), the cases where this resulted in incorrect classification were not numerous[11].

One representative singleton resulting from these errors is exemplified in **D**. *Activate* appears in relatively complicated sentence structures, which gives rise to incorrect SCFs. For example, *MECs cultured on 2D planar substrates transiently* **activate** *MAP kinase in response to EGF, whereas...* gets incorrectly analysed as SCF NP-NP, while *The effect of the constitutively* **activated** *ARF6-Q67L mutant was investigated...* receives the incorrect SCF analysis NP-SCOMP. Most parser errors are caused by unknown domain-specific words and phrases.

## 5 Discussion

Due to differences in the task and experimental setup, direct comparison of our results with previously published ones is impossible. The closest possible comparison point is (Korhonen et al., 2003) which reported 50-59% $m$PUR and 15-19% APP on using IB to assign 110 polysemous (general language) verbs into 34 classes. Our results are substantially better, although we made no effort to restrict our scope to monosemous verbs[12] and although we focussed on a linguistically challenging domain.

It seems that our better result is largely due to the higher uniformity of verb senses in the biomedical domain. We could not investigate this effect systematically because no manually sense annotated data (or a comprehensive list of verb senses) exists for the domain. However, examination of a number of corpus instances suggests that the use of verbs is fairly conventionalized in our data[13]. Where verbs show less sense variation, they show less SCF variation, which aids the discovery of verb classes. Korhonen et al. (2003) observed the opposite with general language data.

We examined, class by class, to what extent our domain-specific gold standard differs from the related general (Levin, 1993) and domain classifications (Spasic et al., 2005; Friedman et al., 2002) (recall that the latter were purely semantic classifications as no lexical ones were available for biomedicine):

33 (of the 50) classes in the gold standard are biomedical. Only 6 of these correspond (fully or mostly) to the semantic classes in the domain classifications. 17 are unrelated to any of the classes in Levin (1993) while 16 bear vague resemblance to them (e.g. our TRANSPORT verbs are also listed under Levin's SEND verbs) but are too different (semantically and syntactically) to be combined.

17 (of the 50) classes are general (scientific) classes. 4 of these are absent in Levin (e.g. QUANTITATE). 13 are included in Levin, but 8 of them have a more restricted sense (and fewer members) than the corresponding Levin class. Only the remaining 5 classes are identical (in terms of members and their properties) to Levin classes.

These results highlight the importance of building or tuning lexical resources specific to different domains, and demonstrate the usefulness of automatic lexical acquisition for this work.

## 6 Conclusion

This paper has shown that current domain-independent NLP and ML technology can be used to automatically induce a relatively high accuracy verb classification from a linguistically challenging corpus of biomedical texts. The lexical classification resulting from our work is strongly domain-specific (it differs substantially from previous ones) and it can be readily used to aid BIONLP. It can provide useful material for investigating the syntax and semantics of verbs in biomedical data or for supplementing existing domain lexical resources with additional information (e.g.

[11] This is partly because the mistakes of the parser are somewhat consistent (similar for similar verbs) and partly because the SCFs gather data from hundreds of corpus instances, many of which are analysed correctly.

[12] Most of our test verbs are polysemous according to WordNet (WN) (Miller, 1990), but this is not a fully reliable indication because WN is not specific to this domain.

[13] The different sub-domains of the biomedical domain may, of course, be even more conventionalized (Friedman et al., 2002).

semantic classifications with additional member verbs). Lexical resources enriched with verb class information can, in turn, better benefit practical tasks such as parsing, predicate-argument identification, event extraction, identification of biomedical relation patterns, among others.

In the future, we plan to improve the accuracy of automatic classification by seeding it with domain-specific information (e.g. using named entity recognition and anaphoric linking techniques similar to those of Vlachos et al. (2006)). We also plan to conduct a bigger experiment with a larger number of verbs and demonstrate the usefulness of the bigger classification for practical BIO-NLP application tasks. In addition, we plan to apply similar technology to other interesting domains (e.g. tourism, law, astronomy). This will not only enable us to experiment with cross-domain lexical class variation but also help to determine whether automatic acquisition techniques benefit, in general, from domain-specific tuning.

## Acknowledgement

We would like to thank Yoko Mizuta, Shoko Kawamato, Sven Demiya, and Parantu Shah for their help in creating the gold standard.

## References

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 $5^{\text{th}}$ *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 $3^{\text{rd}}$ *International Conference on Language Resources and Evaluation*, pages 1499–1504, Las Palmas, Gran Canaria.

A. G. Dimitrov and J. P. Miller. 2001. Neural coding and decoding: communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441–472.

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

C. Friedman, 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.

V. Hatzivassiloglou and W. Weng. 2002. Learning anchor verbs for biological interaction patterns from published text articles. *International Journal of Medical Inf.*, 67:19–32.

L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu. 2002. Accomplishments and challenges in literature data mining for biology. *Journal of Bioinformatics*, 18(12):1553–1561.

T. Hoffman. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

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

A. Korhonen, Y. Krymolowski, and Z. Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Sapporo, Japan.

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

M. Lease and E. Charniak. 2005. Parsing biomedical literature. In *Second International Joint Conference on Natural Language Processing*, pages 58–69.

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.

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

D. Prescher, 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.

I. Spasic, S. Ananiadou, and J. Tsujii. 2005. Masterclass: A case-based reasoning system for the classification of biomedical terms. *Journal of Bioinformatics*, 21(11):2748–2758.

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

A. Vlachos, 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*.