Automatic Lexical Classification - Balancing between Machine Learning and Linguistics*

Anna Korhonen

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

Abstract. Verb classifications have been used to support a number of practical tasks and applications, such as parsing, information extraction, question-answering, and machine translation. However, large-scale exploitation of verb classes in real-world or domain-sensitive tasks has not been possible because existing manually built classifications are incomprehensive. This paper describes recent and on-going research on extending and acquiring lexical classifications automatically. The automatic approach is attractive since it is cost-effective and opens up the opportunity of learning and tuning lexical classifications for the application and domain in question. However, the development of an optimal approach is challenging, and requires not only expertise in machine learning but also a good understanding of the linguistic principles of lexical classification.

Keywords: lexical-semantic classification, verb classes, automatic lexical acquisition

1 Introduction

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

Particularly useful are classes which capture generalizations over a range of (cross-)linguistic properties, such as the ones proposed by Levin (1993). Being defined in terms of similar meaning and (morpho-)syntactic behaviour of words, these classes generally incorporate a wider range of properties than e.g. classes defined solely on semantic grounds (Miller, 1995).

For example, verbs which share the meaning component of 'manner of motion' (e.g. *travel*, *run*, *walk*), behave similarly in terms of subcategorization (e.g. *I travelled/ran/walked*, *I travelled/ran/walked to London*, *I travelled/ran/walked five miles*) and usually have zero-related nominals (e.g. *a run*, *a walk*) can be grouped to the same lexical class. Such verb classes can be identified across the entire lexicon and they can also apply across languages, since the basic meaning components they are comprised of are cross-linguistically applicable or overlapping.

While the classes do not provide means for full semantic inferencing, they can offer a powerful tool for generalization, abstraction and prediction which is beneficial for practical tasks. Fundamentally, the classes are a critical component of any system which needs mapping from surface realization of arguments to predicate-argument structure. As the classes capture higher level abstractions they can be used as a principled means to abstract away from individual words when required. For example, they can be utilized to organize a default inheritance hierarchy which

* This work was funded by the Royal Society University Research Fellowship, the EPSRC grant EP/F030061/1 and the British Council Alliance grant. The author would like to thank Lin Sun for his contribution to this paper.

effectively captures generalizations over words and predicts much of the syntactic/semantic behaviour of a new word simply by associating it with an appropriate class. The predictive power of the classes can help compensate for lack of sufficient data. In addition, the classes have theoretical benefits. For example, classified data can be used to evaluate empirical claims of different linguistic and psycholinguistic theories.

Although lexical classes have proved helpful for a number of (multilingual) tasks, their large-scale exploitation in real-world or highly domain-sensitive tasks has been limited because no fully accurate or comprehensive lexical classification is available. There is no such resource because manual classification of large numbers of words has proved very time-consuming. Class-based differences are typically manifested in differences in the statistics over usages of syntactic-semantic features. This statistical information is difficult collect by hand as it is highly domain-sensitive, i.e. it varies with predominant word senses, which change across corpora and domains.

In recent years, automatic induction of verb classes from corpus data has become increasingly popular (Merlo and Stevenson, 2001; Schulte im Walde, 2006; Joanis *et al.*, 2008; Sun *et al.*, 2008; Li and Brew, 2008; Korhonen *et al.*, 2008; Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009). This work is important as it opens up the opportunity of learning and tuning classifications for the application and domain in question. Automatic classification is not only cost-effective but it also gathers the important statistical information as side effect of the acquisition process and can easily be applied to new domains and usage patterns provided relevant corpus data is available.

To date, a variety of approaches have been proposed for verb classification and applied to general English and other languages. Both supervised and unsupervised machine learning (ML) methods have been used to classify a variety of features extracted from raw, tagged and/or parsed corpus data. Although the results have been generally encouraging, the accuracy of automatic classification shows room for improvement. After providing a short introduction to the basic principles of manual verb classification, this paper reviews recent research in automatic classification – particularly focussing on work conducted in English – and discusses then the various current challenges that need to be met for substantial further advances. Meeting these challenges requires solid expertise in both machine learning and (computational) linguistics.

2 Lexical Classification

The largest and most widely deployed verb classification in English is the classification of Levin (1993). This classification provides a summary of the variety of theoretical research done on lexical-semantic verb classification over the past decades. Verbs which display the same or a similar set of *diathesis alternations* in the realization of their argument structure are assumed to share certain meaning components and are organized into a semantically coherent class. Although alternations are chosen as the primary means for identifying verb classes, additional properties related to subcategorization, morphology and extended meanings of verbs are taken into account as well. For instance, the Levin class of "*Break* Verbs" (class 45.1), which refers to actions that bring about a change in the material integrity of some entity, is characterized by its participation (1-3) or non-participation (4-6) in the following alternations and other constructions (7-8):

1. **Causative/inchoative alternation**:
   Tony broke the window ↔ *The window broke*
2. **Middle alternation**:
   Tony broke the window ↔ *The window broke easily*
3. **Instrument subject alternation**:
   Tony broke the window with the hammer ↔ *The hammer broke the window*
4. ***With/against* alternation**:
   Tony broke the cup against the wall ↔ **Tony broke the wall with the cup*
5. ***Conative alternation**:
   Tony broke the window ↔ **Tony broke at the window*

6. ***Body-Part possessor ascension alternation**:
   **Tony broke herself on the arm* $\leftrightarrow$ *Tony broke her arm*
7. **Unintentional interpretation available (some verbs)**:
   Reflexive object: **Tony broke himself*
   Body-part object: *Tony broke his finger*
8. **Resultative phrase**:
   Tony broke the piggy bank open, Tony broke the glass to pieces

VerbNet Kipper-Schuler (2005)[1] – an extensive on-line lexicon for English verbs – provides detailed syntactic-semantic descriptions of Levin's classes as well as additional classes organized into a refined taxonomy. The resulting taxonomy classifies over 5000 verbs into 274 first level classes. It has been used to support both the development and the evaluation of automatic verb classification.

3 Automatic Verb Classification - the State of the Art

To date, most work on automatic verb classification has focussed on English (Joanis *et al.*, 2008; Sun *et al.*, 2008; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008; Vlachos *et al.*, 2009; Sun and Korhonen, 2009), although some work has also been done on other languages, in particular on German (Schulte im Walde, 2006), and recently also on sub-languages (Korhonen *et al.*, 2008). In this section, we provide an overview of recent work mostly conducted on English (for other languages and domains, see section 5). We will first describe the features and techniques used for classification, and then evaluation and performance of current systems.

3.1 Features

As discussed above in section 2, the main feature of manual verb classification is a diathesis alternation which manifests at the level of syntax in alternating sets of subcategorization frames (SCFs). Since automatic detection of diathesis alternations is challenging (McCarthy, 2001), most work on automatic classification has focussed on syntactic features, exploiting the fact that similar alternations tend to result in similar syntactic behaviour. The syntactic features have been shallow syntactic slots (e.g. NPs preceding or following the verb) extracted using a lemmatizer or a chunker, or verb SCFs extracted using a chunker or a parser. These both feature types have been refined with information about prepositional preferences (PPs) of verbs. Joanis *et al.* (2008) have reported better results using syntactic slots, while several others have obtained good results using SCFs, e.g. (Schulte im Walde, 2006; Li and Brew, 2008; Sun and Korhonen, 2009). While SCFs correspond better (than syntactic slots) with the features used in manual work, optimal results have required including in SCFs also additional information about adjuncts (not only arguments) of verbs (Sun *et al.*, 2008) which are typically not used in manual classification.

Recent research has also experimented with replacing or supplementing SCFs with information about basic lexical context (co-occurrences (COs)) of verbs, or lexical preferences (LPs) in specific grammatical relations (GRs) associated with verbs in parsed data (for example, the type and frequency of prepositions in the indirect object relation) (Li and Brew, 2008; Sun and Korhonen, 2009). Some experiments have also explored the usefulness of verb tense (e.g. the part-of-speech tags of verbs), voice (the knowledge whether the verb was used in active or passive) and/or aspect for verb classification (Joanis *et al.*, 2008; Korhonen *et al.*, 2008).

While most work has focussed on syntactic or lexical features, a few attempts have also been made to refine syntactic features with semantic information about verb selectional preferences (SPs). Following Merlo and Stevenson (2001), Joanis *et al.* (2008) used an simple 'animacy' feature which was determined by classifying e.g. pronouns and proper names in data to this single SP class. Joanis (2002) employed as SP models the top level WordNet (Miller, 1995) classes

[1] See http://verbs.colorado.edu/verb-index/index.php for details.

(Schulte im Walde (2006) tried a similar approach for German). Recently, Sun and Korhonen (2009) experimented with automatically acquired SPs. The latter were obtained by clustering argument head data in GRs related to specific verbs.

Finally, combinations of lexical, syntactic, semantic and other features have been explored.

3.2 Classification

Both supervised and unsupervised machine learning (ML) methods have been used to classify features discussed in the above section. Supervised methods yield optimal performance where adequate and accurate training data are available. A wide range of methods have been employed, including the K Nearest Neighbours, Maximum Entropy, Support Vector Machines, Gaussian, Distributional Kernel methods, and Bayesian Multinomial Regression, among others (Joanis *et al.*, 2008; Sun *et al.*, 2008; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008).

Unsupervised methods have the benefit that they can be used to to discover novel information from corpus data. The latter is particularly useful for supplementing or improving existing classifications or learning new classifications for languages and domains where no manually built classifications are available. Again a range of methods have been explored, including e.g. the K means, Expectation-Maximization, spectral clustering, Information Bottleneck, Probabilistic Latent Semantic Analysis, cost-based pairwise clustering (Brew and Schulte im Walde, 2002; Schulte im Walde, 2006; Korhonen *et al.*, 2008; Sun and Korhonen, 2009; Vlachos *et al.*, 2009). Both soft and hard clustering methods have been tried, but attempts to deal with polysemy (the fact that many verbs can be classified in more than one class) have not been successful yet (see section 4).

3.3 Evaluation

Research on automatic verb classification has typically been evaluated against a manually constructed gold standard. The subsequent sections describe the most commonly used gold standards, evaluation measures, and test sets, and compares the performance of the state-of-the-art approaches for English which have been evaluated using these test sets.

3.3.1 Gold standards The most common evaluation resource in English verb classification has been that of Levin (1993) supplemented with additional information from VerbNet or WordNet. In particular, two gold standards based on (Levin, 1993) have been used to evaluate much of the recent research:

GS1 The gold standard of Joanis *et al.* (2008) provides a classification of 835 verbs into 15 (some coarse, some fine-grained) Levin classes. We consider here the '14 way' version of this resource because this corresponds the closest to the target (Levin's fine-grained) classification[2]. When the frequency-based selection criteria of Joanis *et al.* (2008) is applied and the class imbalance is restricted to 1:1.5, GS1 provides a classification of 205 verbs in 10-15 classes.

GS2 The gold standard of Sun *et al.* (2008) classifies 204 medium-high frequency verbs to 17 fine-grained Levin classes, so that each class has 12 member verbs.

Table 1 from (Sun and Korhonen, 2009) shows the classes in GS1 and GS2.

3.3.2 Data and evaluation measures The classification techniques have been typically applied to large cross-domain corpora and evaluated (against a chosen gold standard) using various measures. Although the measures have differed (e.g. for supervised or unsupervised approaches), the general tendency has been to prefer measures which are (i) applicable to all classification methods under comparison, (ii) deliver a numerical value easy to interpret and (ii) preferably do not introduce biases towards specific numbers of classes or class sizes. The measures mentioned here are

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

Table 1: Levin classes in GS1 and GS2

GS1		GS2	
Object Drop	26.{1,3,7}	Remove	10.1
Recipient	13.{1,3}	Send	11.1
Admire	31.2	Get	13.5.1
Amuse	31.1	Hit	18.1
Run	51.3.2	Amalgamate	22.2
Sound	43.2	Characterize	29.2
Light & Substance	43.{1,4}	Peer	30.3
		Amuse	31.1
Cheat	10.6	Correspond	36.1
Steal & Remove	10.{5,1}	Manner of speaking	37.3
		Say	37.7
Wipe	10.4.{1,2}	Nonverbal expression	40.2
Spray / Load	9.7	Light	43.1
Fill	9.8	Other change of state	45.4
Putting	9.1-6	Mode with motion	47.3
Change of State	45.1-4	Run	51.3.2
		Put	9.1

measures that have been used to evaluate many of the recent clustering approaches compared in the following section:

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

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

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

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

mPUR and ACC have been used as measures of precision (P) and recall(R) respectively. F measure has been calculated as the harmonic mean of P and R:

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

The random baseline (BL) is typically calculated as follows:

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

3.3.3 Performance To give an idea of how current approaches perform, we examined the recent supervised and unsupervised works on general English verb classification which were evaluated on GS1 and GS2 using either the evaluation measures described in the previous section or measures comparable to them. These works are summarized in Table 2. ACC and F-measure are shown for GS1 and GS2, respectively[3].

On GS1[4], the best performing supervised method reported so far is that of Li and Brew (2008). Li and Brew used Bayesian Multinomial Regression for classification. A range of feature sets

[3] A smaller-scale version of this comparison was presented in (Sun and Korhonen, 2009).

[4] Note that the different experiments did not necessarily employ identical sub-sets of GS1 so are not entirely comparable.

Table 2: Performance of recent approaches

		Method	Result
GS1	Li et al. 2008	supervised	66.3
	Joanis et al. 2008	supervised	58.4
	Stevenson et al. 2003	semi-supervised unsupervised	29 31
	Sun and Korhonen 2009	unsupervised	57.55
GS2	Sun et al. 2008	supervised unsupervised	62.50 51.6
	Ó Séaghdha et al. 2008	supervised	67.3
	Sun and Korhonen 2009	unsupervised	80.35

integrating COs, SCFs and/or LPs were extracted from a large corpus using a lemmatizer and a grammatical parser. The combination of COs and SCFs gave the best result, shown in the table.

Joanis *et al.* (2008) have reported the second best supervised result on GS1, using Support Vector Machines for classification. They compared various features derived from linguistic analysis and extracted using shallow syntactic processing (mainly chunking): syntactic slots, slot overlaps, tense, voice, aspect, and animacy of NPs. They concluded that syntactic information about core constituents occurring with a verb (syntactic slots) is most important to verb classification. Stevenson and Joanis (2003) reached a similar conclusion in their unsupervised experiment on GS1. A feature set similar to that of Joanis *et al.* (2008) was employed (features were selected in a semi-supervised fashion) and hierarchical clustering was used.

The recent unsupervised method of Sun and Korhonen (2009) performs better on GS1 than the unsupervised method of Joanis *et al.* (2008) and nearly as well as the supervised approach of Joanis *et al.* (2008). Sun and Korhonen used a variation of spectral clustering based on the MNCut algorithm (Meila and Shi, 2001) and experimented with a variety of features (e.g. COs, SCFs, LPs, voice, tense), including also semantic ones (SPs). The features were extracted using a SCF acquisition system which makes use of a grammatical parser. The SPs were obtained by clustering argument head data in relevant syntactic slots. The best result was obtained when using SCFs in conjunction with SPs.

On GS2, the best performing supervised method so far is that of Ó Séaghdha and Copestake (2008) which employs a distributional kernel method to classify SCF features parameterized for prepositions in the automatically acquired VALEX SCF lexicon. Using exactly the same data and feature set, Sun *et al.* (2008) obtained a slightly lower result when using a supervised method (Gaussian) and a notably lower result when using an unsupervised method (pairwise clustering). The recent unsupervised approach of Sun and Korhonen (2009) (discussed above with GS1) outperforms these both methods on this gold standard when SCFs are used in conjunction with automatically acquired SPs.

Although this brief comparison focuses on recent work on English classification and does not cover approaches evaluated on other gold standards, languages or domains, it does serve to summarise the state of the art: current approaches perform at their best around 66 accuracy and 80 F measure. While this performance is clearly better than the baseline (chance) performance on the task and is likely to be high enough to benefit many practical tasks, it is still much lower than the realistic upper bound for the task: Merlo and Stevenson (2001) estimated that the accuracy of classification performed by experts in lexical classification is likely to be around 85%.

4 Current Challenges

This section discusses the various challenges that need to be met in order to improve the state of the art further.

4.1 Features

Section 3.1 reviewed the features employed in verb classification so far. Section 3.3.3 showed that to date, syntactic features (syntactic slots and SCFs) have been the most useful features in verb classification. Although semantic features play a key role in manual verb classification and could thus be expected to offer a considerable contribution to automatic classification, they have not proved equally successful. Until recently, no significant additional improvement was reported using verb SPs (Joanis, 2002; Schulte im Walde, 2006). This was surprising since SPs are strong indicators of diathesis alternations (McCarthy, 2001) and fairly precise semantic descriptions can be assigned to the majority of Levin classes (Kipper-Schuler, 2005). However, in their recent experiment, Sun and Korhonen (2009) obtained a considerable improvement using SPs in conjunction with syntactic features on both GS1 and GS2, although they used a fully unsupervised approach to both verb clustering and SP acquisition. This suggests that NLP and ML techniques have now developed to the point where the use of deeper, theoretically-motivated features is becoming feasible. Yet high accuracy SP acquisition from undisambiguated corpus data is still an unmet challenge and is especially complex in the context of verb classification where SP models are needed for specific syntactic slots for which the data may be sparse. Recently a number of techniques have been proposed which may offer ideas for further improvement of the approach (Erk, 2007; Bergsma *et al.*, 2008; Schulte im Walde *et al.*, 2008). The number and type (and combination) of GRs for which SPs can be reliably acquired, especially when the data is sparse, requires also further investigation.

However, the main semantic features in manual classification are actually diathesis alternations. Some studies have attempted automatic alternation detection using WordNet classes as SP models (Lapata, 1999; McCarthy, 2001), but no recent large-scale work has been conducted, and no attepts have been made to detect alternations in a fully unsupervised fashion. The time may now be ripe for this research and its integration in verb classification. The development of an optimal approach will require a good understanding of the linguistic basis of verb classification as well as adequate NLP and ML expertise. The approach will need to be general enough to cover most types of alternations, efficient enough for a large scale use and resistant to the problems of sparse data.

4.2 Classification

In section 3.2 we reviewed various supervised and unsupervised methods that have been used for automatic classification. For optimal results, the choice of a machine learning method is not random but involves understanding of the basic principles of the method and its suitability for the data and the task. For example, Sun and Korhonen (2009) obtained promising results in their recent experiment with SP features not only because the features made theoretical sense but also because the clustering method (spectral clustering) was particularly suited for the resulting, high dimensional feature space. Novel ML methods have been developed recently which combine clustering with an element of guidance based on a prior intuition and have useful properties such as not having to define the number of clusters in advance (e.g. unsupervised and constrained Dirichlet Process Mixture Models for verb clustering by Vlachos *et al.* (2009)). This shows the benefit of following the recent developments in the ML community. However, semi-supervised approaches have not been used for the task yet (except for the sub-task of feature selection) although they are well-known in the NLP community and would combine the benefits of supervised and unsupervised approaches (Abney, 2008).

4.3 Polysemy

Polysemy is frequent in language. In particular, many high frequency verbs have several senses and can therefore be members of several classes. Most work on automatic classification has bypassed this issue by assuming a single class for each verb – usually the one corresponding to its predominating (the most frequent sense) in language according to e.g. WordNet. This is not only unrealistic thinking of real-world application of verb classes but also the predominating sense is not static but varies across domains and sub-languages.

Few attempts have been made to address this problem. Korhonen *et al.* (2003) performed a clustering experiment with highly polysemous verbs. They constructed a polysemous gold standard for c. 200 English verbs and examined whether a soft clustering method (Information Bottleneck) could be used to assign these verbs to several classes. The clustering turned out hard, with the majority of verbs being assigned to one class only. Yet the investigation showed that polysemy has a considerable impact on verb classification: optimal results were obtained with when clustering was evaluated against the polysemous gold standard, not the monosemous version of it which assumed the predominant sense according to WordNet.

Clearly polysemy is an issue that needs to be dealt with, and this amounts to both extending gold standards to capture non-predominant senses as well as finding a suitable ML method. Recently a multi-label classification method was used for supervised adjective classification Boleda *et al.* (2007) which might yield useful results also with verbs. Also methods for modelling the overlap between lexical categories might be of use.

4.4 Other languages and domains

Most work on verb classification has been conducted on English. Considerable research has also been done on German (Schulte im Walde, 2006), but only small scale experiments exist on other languages, e.g. Chinese, Italian (Merlo *et al.*, 2002), Spanish (Ferrer, 2004) and Japanese (Oishi and Matsumoto, 1997). Evaluating the applicability of classification techniques to several languages is critical for both theoretical and practical reasons; for 1) improving the accuracy, scalability and robustness of the techniques mainly developed for English or German, 2) advancing work in other languages, 3) gaining a better understanding of the language-specific / cross-linguistic components of lexical information (e.g. the extent to which the features used for English or German are also valid for other languages), and 4) in a long term, improving the performance of such multilingual NLP applications (e.g. machine translation, information extraction) which can benefit from verb classes.

The same can be said also about different domains and sub-languages. The only work (which we are aware of) which has applied verb classification technology to a specific domain is that of Korhonen *et al.* (2008). This work focussed on the important domain of biomedicine for which no large verb classification was available. It involved learning a classification using clustering technology originally developed for general English. The experiment revealed interesting facts about automatic classification, e.g. the fact that domain-specific classifications can be very different from general classifications (even the shared verb classes may have a specialised, narrower meaning). Also, the features performed differently than in general language classification. The fact that many domains tend to be more uniform or conventionalized in terms of language use has many consequences for automatic classification which require further investigation.

4.5 Evaluation

Most evaluation has been quantitative in nature and involved the gold standards discussed earlier in section 3.3.1. While these gold standards provide suitably small test sets for thorough evaluation, it would be important to also investigate the extent to which existing approaches generalise across the entire language. Whilst the classification of over 5000 word senses offered by VerbNet may

not be fully comprehensive, it does offer a valuable larger resource for evaluation.

For many languages and domains, no evaluation resources are available. Both manual (Kipper *et al.*, 2008) and semi-automatic methods (Korhonen *et al.*, 2008) have been proposed for building gold standards from scratch. For example, in the recent work on biomedical verb classification, human experts (linguists and biologists) constructed a gold standard by examining verb classes formed on the basis on syntactic similarity and deciding which ones of them were also semantically related (Korhonen *et al.*, 2008). However, such work requires not only clear guidelines but also adequate linguistic and/or domain expertise.

Some of the works have supplemented quantitative evaluation with qualitative analysis. This has required also linguistic (or domain) expertise, and interestingly, has not only helped to find error types but has often also shown that automatic classification can discover novel, valuable information in data, e.g. classes which are actually related although distinct in a gold standard or classes which are distinct in a gold standard although ought to be related (Schulte im Walde, 2006; Sun *et al.*, 2008; Korhonen *et al.*, 2008; Vlachos *et al.*, 2009). Qualitative evaluation can thus show the true potential of automatic classification and is therefore vital for further development of classification technology. Equally important is evaluation in the context of practical tasks and applications. To the best of our knowledge, no approaches to automatic verb classification have been evaluated in this manner, although the work on automatic verb classification is largely motivated by the practical potential of accurate and relevant classifications.

5 Conclusion

During the past years, a lot has been achieved in automatic verb classification. Yet a lot remains to be done in terms of improving and extending current technology and applying it to larger data sets and novel (sub-)languages. This paper has discussed the various areas which require further improvement (ranging from features to evaluation techniques) and highlighted the fact that further improvements can only be obtained by combining the best available (computational) linguistic and ML expertise.

References

Abney, S. 2008. *Semisupervised Learning for Computational Linguistics*. Chapman and Hall.

Bergsma, S., D. Lin, and R. Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proc. of EMNLP*.

Boleda, G., S. Schulte im Walde, and T. Badia. 2007. Modelling polysemy in adjective classes by multi-label classification. In *Proc. of EMNLP-CoNLL*.

Brew, C. and S. Schulte im Walde. 2002. Spectral clustering for German verbs. In *Proc. of EMNLP*.

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

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

Ferrer, E. E. 2004. Towards a semantic classification of spanish verbs based on subcategorisation information. In *Proceedings of the ACL 2004 Workshop on Student Research*.

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

Joanis, E., S. Stevenson, and D. James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*.

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

Korhonen, A., Y. Krymolowski, and N. Collier. 2008. The Choice of Features for Classification of Verbs in Biomedical Texts. In *Proc. of COLING*.

Korhonen, A., Y. Krymolowski, and Z. Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proc. of ACL*, pages 64–71.

Lapata, M. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proc. of ACL*.

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

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

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

Meila, M. and J. Shi. 2001. A random walks view of spectral segmentation. AISTATS.

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

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

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

Oishi, A. and Y. Matsumoto. 1997. Detecting the organization of semantic subclasses of Japanese verbs. In *International Journal of Corpus Linguistics*, volume 2, pages 65–89.

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

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

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

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

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

Sun, L. and A. Korhonen. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. In *Proc. of EMNLP*.

Sun, L., A. Korhonen, and Y. Krymolowski. 2008. Verb class discovery from rich syntactic data. *Lecture Notes in Computer Science*, 4919:16.

Swier, R. and S. Stevenson. 2004. Unsupervised semantic role labelling. In *Proc. of EMNLP*.

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

Zapirain, B., E. Agirre, and L. Màrquez. 2008. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proc. of ACL*.