Diathesis alternation approximation for verb clustering

Lin Sun
Greedy Intelligence Ltd
Hangzhou, China
lin.sun@greedyint.com

Diana McCarthy and Anna Korhonen
DTAL and Computer Laboratory
University of Cambridge
Cambridge, UK
diana@dianamccarthy.co.uk
alk23@cam.ac.uk

Abstract

Although diathesis alternations have been used as features for manual verb classification, and there is recent work on incorporating such features in computational models of human language acquisition, work on large scale verb classification has yet to examine the potential for using diathesis alternations as input features to the clustering process. This paper proposes a method for approximating diathesis alternation behaviour in corpus data and shows, using a state-of-the-art verb clustering system, that features based on alternation approximation outperform those based on independent subcategorization frames. Our alternation-based approach is particularly adept at leveraging information from less frequent data.

1 Introduction

Diathesis alternations (DAs) are regular alternations of the syntactic expression of verbal arguments, sometimes accompanied by a change in meaning. For example, *The man broke the window* ↔ *The window broke*. The syntactic phenomena are triggered by the underlying semantics of the participating verbs. Levin (1993)'s seminal book provides a manual inventory both of DAs and verb classes where membership is determined according to participation in these alternations. For example, most of the COOK verbs (e.g. bake, cook, fry ...) can all take various DAs, such as the causative alternation, middle alternation and instrument subject alternation.

In computational linguistics, work inspired by Levin's classification has exploited the link between syntax and semantics for producing classifications of verbs. Such classifications are useful for a wide variety of purposes such as semantic role labelling (Gildea and Jurafsky, 2002), predicting unseen syntax (Parisien and Stevenson, 2010), argument zoning (Guo *et al.*, 2011) and metaphor identification (Shutova *et al.*, 2010). While Levin's classification can be extended manually (Kipper-Schuler, 2005), a large body of research has developed methods for automatic verb classification since such methods can be applied easily to other domains and languages.

Existing work on automatic classification relies largely on syntactic features such as subcategorization frames (SCF)s (Schulte im Walde, 2006; Sun and Korhonen, 2011; Vlachos *et al.*, 2009; Brew and Schulte im Walde, 2002). There has also been some success incorporating selectional preferences (Sun and Korhonen, 2009).

Few have attempted to use, or approximate, diathesis features directly for verb classification although manual classifications have relied on them heavily, and there has been related work on identifying the DAs themselves automatically using SCF and semantic information (Resnik, 1993; McCarthy and Korhonen, 1998; Lapata, 1999; McCarthy, 2000; Tsang and Stevenson, 2004). Exceptions to this include Merlo and Stevenson (2001), Joanis *et al.* (2008) and Parisien and Stevenson (2010, 2011). Merlo and Stevenson (2001) used cues such as passive voice, animacy and syntactic frames coupled with the overlap of lexical fillers between the alternating slots to predict a 3-way classification (unergative, unaccusative and object-drop). Joanis *et al.* (2008) used similar features to classify verbs on a much larger scale. They classify up to 496 verbs using 11 different classifications each having between 2 and 14 classes. Parisien and Stevenson (2010, 2011) used hierarchical Bayesian models on slot frequency data obtained from child-directed speech parsed with a dependency parser to model acquisition of SCF, alternations and ultimately verb classes which provided predictions for unseen syntactic behaviour of class members.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 736–741,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

Frame	Example sentence	Freq
NP+PPon	Jessica sprayed paint on the wall	40
NP+PPwith	Jessica sprayed the wall with paint	30
PPwith	*The wall sprayed with paint	0
PPon	Jessica sprayed paint on the wall	30

Table 1: Example frames for verb spray

In this paper, like Sun and Korhonen (2009); Joanis *et al.* (2008) we seek to automatically classify verbs into a broad range of classes. Like Joanis et al., we include evidence of DA, but we do not manually select features attributed to specific alternations but rather experiment with syntactic evidence for alternation approximation. We use the verb clustering system presented in Sun and Korhonen (2009) because it achieves state-of-the-art results on several datasets, including those of Joanis et al., even without the additional boost in performance from the selectional preference data. We are interested in the improvement that can be achieved to verb clustering using approximations for DAs, rather than the DA per se. As such we make the simple assumption that if a pair of SCFs tends to occur with the same verbs, we have a potential occurrence of DA. Although this approximation can give rise to false positives (pairs of frames that co-occur frequently but are not DA) we are nevertheless interested in investigating its potential usefulness for verb classification. One attractive aspect of this method is that it does not require a pre-defined list of possible alternations.

2 Diathesis Alternation Approximation

A DA can be approximated by a pair of SCFs. We parameterize frames involving prepositional phrases with the preposition. Example SCFs for the verb "spray" are shown in Table 1. The feature value of a single frame feature is the frequency of the SCF. Given two frames $f_v(i), f_v(j$ of a verb v, they can be transformed into a feature pair $(f_v(i), f_v(j))$ as an approximation to a DA. The feature value of the DA feature $(f_v(i), f_v(j))$ is approximated by the joint probability of the pair of frames $p(f_v(i), f_v(j)|v)$, obtained by integrating all the possible DAs. The key assumption is that the joint probability of two SCFs has a strong correlation with a DA on the grounds that the DA gives rise to both SCFs in the pair. We use the DA feature $(f_v(i), f_v(j))$ with its value $p(f_v(i), f_v(j)|v)$ as a new feature for verb clustering. As a comparison point, we can ignore the DA and make a frame independence assumption. The joint probability is decomposed as:

$$p(f_v(i), f_v(j)|v)' \triangleq p(f_v(i)|v) \cdot p(f_v(j)|v) \quad (1)$$

We assume that SCFs are dependent as they are generated by the underlying meaning components (Levin and Hovav, 2006). The frame dependency is represented by a simple graphical model in figure 1.

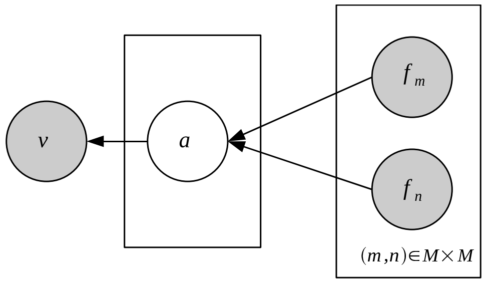

Figure 1: Graphical model for the joint probability of pairs of frames. v represents a verb, a represents a DA and f represents a specific frame in total of M possible frames

In the data, the verb (v) and frames (f) are observed, and any underlying alternation (a) is hidden. The aim is to approximate but not to detect a DA, so a is summed out:

$$p(f_v(i), f_v(j)|v) = \sum_a p(f_v(i), f_v(j)|a) \cdot p(a|v) \quad (2)$$

In order to evaluate this sum, we use a relaxation [1]: the *sum* in equation 1 is replaced with the maximum (*max*). This is a reasonable relaxation, as a pair of frames rarely participates in more than one type of a DA.

$$p(f_v(i), f_v(j)|v) \approx \max(p(f_v(i), f_v(j)|a) \cdot p(a|v)) \quad (3)$$

The second relaxation further relaxes the first one by replacing the *max* with the least upper bound (*sup*): If $f_v(i)$ occurs a times, $f_v(j)$ occurs b times and $b < a$, the number of times that a DA occurs between $f_v(i)$ and $f_v(j)$ must be smaller or equal to b.

$$p(f_v(i), f_v(j)|v) \approx \sup\{p(f_v(i), f_v(j)|a)\} \cdot \sup\{p(a|v)\} \quad (4)$$

$$\begin{aligned} \sup\{p(f_v(i), f_v(j)|a)\} &= Z^{-1} \cdot \min(f_v(i), f_v(j)) \\ \sup\{p(a|v)\} &= 1 \\ Z &= \sum_m \sum_n \min(f_v(m), f_v(n)) \end{aligned}$$

[1] A relaxation is used in mathematical optimization for relaxing the strict requirement, by either substituting it with an easier requirement or dropping it completely.

Frame pair	Possible DA	Frequency
NP+PPon NP+PPwith	Locative	30
NP+PPon PPwith	Causative(with)	0
NP+PPon PPon	Causative(on)	30
NP+PPwith PPwith	?	0
NP+PPwith PPon	?	30
PPwith PPon	?	0
NP+PPon NP+PPon	-	40
NP+PPwith NP+PPwith	-	30
PPwith PPwith	-	0
PPon PPon	-	30

Table 2: Example frame pair features for *spray*

So we end up with a simple form:

$$p(f_v(i), f_v(j)|v) \approx Z^{-1} \cdot \min(f_v(i), f_v(j)) \quad (5)$$

The equation is intuitive: If $f_v(i)$ occurs 40 times and $f_v(j)$ 30 times, the DA between $f_v(i)$ and $f_v(j) \leq 30$ times. This upper bound value is used as the feature value of the DA feature. The original feature vector $\mathbf{f}$ of dimension M is transformed into M^2 dimensions feature vector $\tilde{\mathbf{f}}$. Table 2 shows the transformed feature space for *spray*. The feature space matches our expectation well: valid DAs have a value greater than 0 and invalid DAs have a value of 0.

3 Experiments

We evaluated this model by performing verb clustering experiments using three feature sets:

F1: SCF parameterized with preposition. Examples are shown in Table 1.

F2: The frame pair features built from F1 with the frame independence assumption (equation 1). This feature is not a DA feature as it ignores the inter-dependency of the frames.

F3: The frame pair features (DAs) built from F1 with the frame dependency assumption (equation 4). This is the DA feature which considers the correlation of the two frames which are generated from the alternation.

F3 implicitly includes F1, as a frame can pair with itself. [2] In the example in Table 2, the frame pair "PP(on) PP(on)" will always have the same value as the "PP(on)" frame in F1.

We extracted the SCFs using the system of Preiss *et al.* (2007) which classifies each corpus occurrence of a verb as a member of one of the 168 SCFs on the basis of grammatical relations identified by the RASP (Briscoe *et al.*, 2006) parser. We experimented with two datasets that have been used in prior work on verb clustering: the test sets 7-11 (3-14 classes) in Joanis *et al.* (2008), and the 17 classes set in Sun *et al.* (2008).

We used the spectral clustering (SPEC) method and settings as in Sun and Korhonen (2009) but adopted the Bhattacharyya kernel (Jebara and Kondor, 2003) to improve the computational efficiency of the approach given the high dimensionality of the quadratic feature space.

$$w_b(v, v') = \sum_{d=1}^{D} (v_d v'_d)^{1/2} \qquad (6)$$

The mean-filed bound of the Bhattacharyya kernel is very similar to the KL divergence kernel (Jebara *et al.*, 2004) which is frequently used in verb clustering experiments (Korhonen *et al.*, 2003; Sun and Korhonen, 2009).

To further reduce computational complexity, we restricted our scope to the more frequent features. In the experiment described in this section we used the 50 most frequent features for the 3-6 way classifications (Joanis et al.'s test set 7-9) and 100 features for the 7-17 way classifications. In the next section, we will demonstrate that F3 outperforms F1 regardless of the feature number setting. The features are normalized to sum 1.

The clustering results are evaluated using F-Measure as in Sun and Korhonen (2009) which provides the harmonic mean of precision (P) and recall (R)

P is calculated using modified purity – a global measure which evaluates the mean precision of clusters. Each cluster ($k_i \in K$) is associated with the gold-standard class to which the majority of its members belong. The number of verbs in a cluster (k_i) that take this class is denoted by $n_{prevalent}(k_i)$.

$$P = \frac{\sum_{k_i \in K: n_{prevalent(k_i)} > 2} n_{prevalent(k_i)}}{|\text{verbs}|}$$

R is calculated using weighted class accuracy: the proportion of members of the dominant cluster DOM-CLUST$_i$ within each of the gold-standard classes $c_i \in C$.

[2] We did this so that F3 included the SCF features as well as the DA approximation features. It would be possible in future work to exclude the pairs involving identical frames, thereby relying solely on the DA approximations, and compare performance with the results obtained here.

	Datasets					
	Joanis et al.					Sun et al.
	7	8	9	10	11	
F1	54.54	49.97	35.77	46.61	38.81	60.03
F2	50.00	49.50	32.79	54.13	40.61	64.00
F3	**56.36**	**53.79**	**52.90**	**66.32**	**50.97**	**69.62**

Table 3: Results when using F3 (DA), F2 (pair of independent frames) and F1 (single frame) features with Bhattacharyya kernel on Joanis et al. and Sun et al. datasets

$$R = \frac{\sum_{i=1}^{|C|} |\text{verbs in DOM-CLUST}_i|}{|\text{verbs}|}$$

The results are shown in Table 3. The result of F2 is lower than that of F3, and even lower than that of F1 for 3-6 way classification. This indicates that the frame independence assumption is a poor assumption. F3 yields substantially better result than F2 and F1. The result of F3 is 6.4% higher than the result (F=63.28) reported in Sun and Korhonen (2009) using the F1 feature.

This experiment shows, on two datasets, that DA features are clearly more effective than the frame features for verb clustering, even when relaxations are used.

4 Analysis of Feature Frequency

A further experiment was carried out using F1 and F3 on Joanis *et al.* (2008)'s test sets 10 and 11. The frequency ranked features were added to the clustering one at a time, starting from the most frequent one. The results are shown in figure 2. F3 outperforms F1 clearly on all the feature number settings. After adding some highly frequent frames (22 for test set 10 and 67 for test set 11), the performance for F1 is not further improved. The performance of F3, in contrast, is improved for almost all (including the mid-range frequency) frames, although to a lesser degree for low frequency frames.

5 Related work

Parisien and Stevenson (2010) introduced a hierarchical Bayesian model capable of learning verb alternations and constructions from syntactic input. The focus was on modelling and explaining the child alternation acquisition rather than on automatic verb classification. Therefore, no quantitative evaluation of the clustering is reported, and the number of verbs under the novel verb generalization test is relatively small. Parisien and Stevenson (2011) extended this work by adding semantic features.

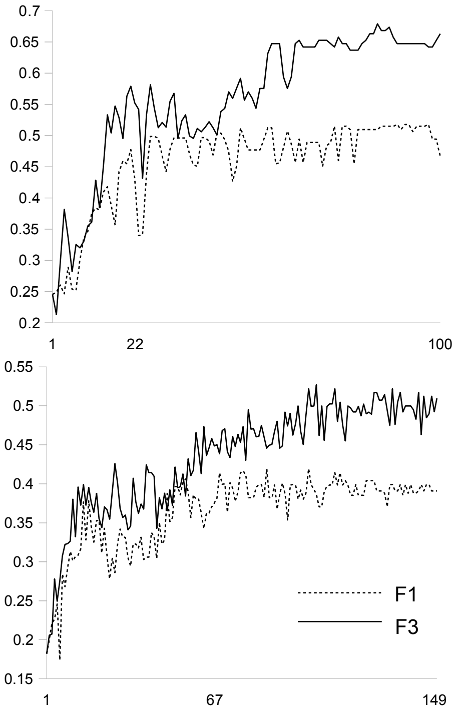

Figure 2: Comparison between frame features (F1) and DA features (F3) with different feature number settings. DA features clearly outperform frame features. The top figure is the result on test set 10 (8 ways). The bottom figure is the result on test set 11 (14 ways). The x axis is the number of features. The y axis is the F-Measure result.

Parisien and Stevenson's (2010) model 2 has a similar structure to the graphic model in figure 1. A fundamental difference is that we explicitly use a probability distribution over alternations (pair of frames) to represent a verb, whereas they represent a verb by a distribution over the observed frames similar to Vlachos *et al.* (2009) 's approach. Also the parameters in their model were inferred by Gibbs sampling whereas we avoided this inference step by using relaxation.

6 Conclusion and Future work

We have demonstrated the merits of using DAs for verb clustering compared to the SCF data from which they are derived on standard verb classification datasets and when integrated in a state-of-the-art verb clustering system. We have also demonstrated that the performance of frame features is dominated by the high frequency frames. In contrast, the DA features enable the mid-range frequency frames to further improve the performance.

In the future, we plan to evaluate the performance of DA features in a larger scale experiment. Due to the high dimensionality of the transformed feature space (quadratic of the original feature space), we will need to improve the computational efficiency further, e.g. via use of an unsupervised dimensionality reduction technique Zhao and Liu (2007). Moreover, we plan to use Bayesian inference as in Vlachos *et al.* (2009); Parisien and Stevenson (2010, 2011) to infer the actual parameter values and avoid the relaxation.

Finally, we plan to supplement the DA feature with evidence from the slot fillers of the alternating slots, in the spirit of earlier work (McCarthy, 2000; Merlo and Stevenson, 2001; Joanis *et al.*, 2008). Unlike these previous works, we will use selectional preferences to generalize the argument heads but will do so using preferences from distributional data (Sun and Korhonen, 2009) rather than WordNet, and use *all argument head data* in *all* frames. We envisage using maximum average distributional similarity of the argument heads in any potentially alternating slots in a pair of co-occurring frames as a feature, just as we currently use the frequency of the less frequent co-occurring frame.

Acknowledgement

Our work was funded by the Royal Society University Research Fellowship (AK) and the Dorothy Hodgkin Postgraduate Award (LS).

References

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

E. Briscoe, J. Carroll, and R. Watson. The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80, 2006.

D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

Y. Guo, A. Korhonen, and T. Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of EMNLP*, pages 273–283, Stroudsburg, PA, USA, 2011. ACL.

T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop*, page 57. Springer, 2003.

T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *The Journal of Machine Learning Research*, 5:819–844, 2004.

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

K. Kipper-Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA, June 2005.

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

M. Lapata. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of ACL*, pages 397–404. ACL Morristown, NJ, USA, 1999.

B. Levin and M. Hovav. Argument realization. *Computational Linguistics*, 32(3):447–450, 2006.

B. Levin. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London, 1993.

D. McCarthy and A. Korhonen. Detecting verbal participation in diathesis alternations. In *Proceedings of ACL*, volume 36, pages 1493–1495. ACL, 1998.

D. McCarthy. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of NAACL*, pages 256–263. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2000.

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

C. Parisien and S. Stevenson. Learning verb alternations in a usage-based Bayesian model. In *Proceedings of the 32nd annual meeting of the Cognitive Science Society*, 2010.

C. Parisien and S. Stevenson. Generalizing between form and meaning using learned verb classes. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 2011.

J. Preiss, T. Briscoe, and A. Korhonen. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL*, volume 45, page 912, 2007.

P. Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, 1993.

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

E. Shutova, L. Sun, and A. Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of COLING*, pages 1002–1010. ACL, 2010.

L. Sun and A. Korhonen. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of EMNLP*, pages 638–647, 2009.

L. Sun and A. Korhonen. Hierarchical verb clustering using graph factorization. In *Proceedings of EMNLP*, pages 1023–1033, Edinburgh, Scotland, UK., July 2011. ACL.

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

V. Tsang and S. Stevenson. Using selectional profile distance to detect verb alternations. In *HLT/NAACL 2004 Workshop on Computational Lexical Semantics*, 2004.

A. Vlachos, A. Korhonen, and Z. Ghahramani. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 74–82, 2009.

Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of ICML*, pages 1151–1157, New York, NY, USA, 2007. ACM.