In Proceedings of the Fifth Workshop on Computational Natural Language Learning (CoNLL-2001), 30-37.

Automatic Verb Classification Using Multilingual Resources

Vivian Tsang
Department of Computer Science
University of Toronto
10 King's College Road
Toronto, ON
Canada M5S 3G4
vyctsang@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
6 King's College Road
Toronto, ON
Canada M5S 3H5
suzanne@cs.toronto.edu

Abstract

We propose the use of multilingual corpora in the automatic classification of verbs. We extend the work of (Merlo and Stevenson, 2001), in which statistics over simple syntactic features extracted from textual corpora were used to train an automatic classifier for three lexical semantic classes of English verbs. We hypothesize that some lexical semantic features that are difficult to detect superficially in English may manifest themselves as easily extractable surface syntactic features in another language. Our experimental results combining English and Chinese features show that a small bilingual corpus may provide a useful alternative to using a large monolingual corpus for verb classification.

1 Introduction

Recently, a number of researchers have devised corpus-based approaches for automatically learning the lexical semantic class of verbs (e.g., (McCarthy and Korhonen, 1998; Lapata and Brew, 1999; Schulte im Walde, 2000; Merlo and Stevenson, 2001)). Automatic verb classification yields important potential benefits for the creation of lexical resources. Lexical semantic classes incorporate both syntactic and semantic information about verbs, such as the general sense of the verb (e.g., change-of-state or manner-of-motion) and the allowable mapping of verbal arguments to syntactic positions (e.g., whether an experiencer argument can appear as the subject or the object of the verb) (Levin, 1993). By automatically learning the assignment of verbs to lexical semantic classes, each verb inherits a great deal of information about its possible usage in an NLP system, without that information having to be explicitly hand-coded.

In this paper, we explore the use of multilingual corpora in the automatic learning of verb classification. We extend the work of (Merlo and Stevenson, 2001), in which statistics over simple syntactic features extracted from syntactically annotated corpora were used to train an automatic classifier for a set of sample lexical semantic classes of English verbs. This work had two potential limitations: first, only a small number (five) of syntactic features that correlate with semantic class were proposed; second, a very large corpus was needed (65M words) to extract sufficiently discriminating statistics.

We address both of these issues in the current study by exploiting the use of a parallel English-Chinese corpus. Our motivating hypothesis is that some lexical semantic features that are difficult to detect superficially in English may manifest themselves as surface syntactic features in another language. If this is indeed the case, then we should be able to augment the initial set of English features with features over the translated verbs in the other language (in our case, Chinese).

Our hypothesis that a non-English verb feature set can be useful in English verb classification is inspired by SLA (Second Language Acquisition) research on learning English verbs. As the name suggests, SLA research studies how humans acquire a second language. "Transfer effects"—the impact of one's native language when learning a second language (Ellis, 1997)—are of particular interest to us. Recent research has shown that properties of a non-English native lexicon can influence human learning of English verb class distinctions (e.g., (Helms-Park, 1997; Inagaki, 1997; Juffs, 2000)). Carrying this idea of "transfer" over to the machine learning setting, we hypothesize that features from a second language may provide an additional source of information that complements the English features, making it possible that a smaller corpus (a bitext) can be a useful alternative to using a large monolingual corpus for verb classification.

2 The Verb Classes and English Features

Merlo and Stevenson (2001) tested their approach on the major classes of optionally intransitive verbs in English. All the classes allow the same subcategorizations (transitive and intransitive), entailing that they cannot be discriminated by subcategorization alone. Thus, successful classification demonstrates the induction of semantic information from syntactic features.

In our work, we focus on two of these classes, the change-of-state verbs, such as *open*, and the verbs of creation and transformation, such as *perform* (classes 45 and 26, respectively, from (Levin, 1993)). Both classes are optionally intransitive, but differ in the alternation between the transitive and intransitive forms. The transitive form of a change-of-state verb is a causative form of the intransitive (*the door opened/the cat opened the door*), while the transitive/intransitive alternates of a creation/transformation verb arise from simple object optionality (*the actors performed the skit/the actors performed*).

Merlo and Stevenson (2001) used 5 numeric features that encoded summary statistics over the usage of each verb across the corpus (65M words of Wall Street Journal, WSJ). The features captured subcategorization and aspectual frequencies (of transitivity, passive voice, and VBN POS tag), as well as statistics that approximated thematic properties of NP arguments (animacy and causativity) from simple syntactic indicators. We adopt these same features in our work, and augment them with Chinese features as described next.

3 Chinese Features

We selected the following Chinese features for our task, based on the properties of the change-of-state and creation/transformation classes. Each numbered item refers to a collection of related features. We describe how we expect each type of feature to vary across the two classes.

1. **Chinese POS tags for Verbs**: We used the CKIP (Chinese Knowledge Information Processing Group) POS-tagger to assign one of 15 verb tags to each verb. Additionally, each of these tags can be mapped into the UPenn Chinese Treebank standard (Fei Xia, email communication), which characterizes each verb as "active" or "stative".

   We note that change-of-state verbs are more likely to be adjectivized than creation/transformation verbs; furthermore, this adjectival property is not unlike the stative property in Chinese. We expect then to see the Chinese translation of English change-of-state verbs to be more likely assigned a stative verb tag.

2. **Passive Particles**: The adjectival nature of change-of-state verbs may also be reflected in a higher proportion of passive use, since the adjectival use is a passive use. In Chinese, a passive construction is indicated by a passive particle preceding the main verb. For example, the passive sentence:

   This store is closed.

   can be translated as:

   Zhe4 ge4 (this) shang1 dian4 (store) bei4 (passive particle) guan1 bi4 (closed).

   We thus expect to find that translations of change-of-state verbs have a higher frequency of occurrence with a passive particle in Chinese.

3. **Periphrastic (Causative) Particles**: In Chinese, some causative sentences use an external (periphrastic) particle to indicate that the subject is the causal agent of the event specified by the verb. For example, one possible translation for

   I cracked an egg.

   can be

   Wo3 (I) jiang1 (made, periphrastic particle) dan4 (egg) da3 lan4 (crack).

   Since change-of-state verbs have a causative alternate, and creation/transformation verbs do not, we expect to see a more frequent use of such particles in the translated equivalent of the change-of-state verbs.

4. **Morpheme Information**: The types of features discussed so far involve the POS tag of the translated verb, or additional syntactic particles it occurs with. We also hypothesize that the semantic class membership of an English verb may influence its word-level translation into Chinese. That is, the sublexical component—the precise morphemic constitution of the translated Chinese verb—may reflect properties of the class of the English verb. The following features are an attempt to exploit this potential source of information:

   - Average number of morphemes in translated verb.
   - Different categories of morphemes in translated verb. (We count occurrences of all combinations of pairs of POS tags V, N, and A.)
   - Semantic specificity of translated verb. (Is it semantically more specific than the English verb, e.g., by including additional morphemes?)

The four general types of features we describe above lead to 17 Chinese features in total, which we use alone or in combination with the original 5 features proposed by Merlo and Stevenson (2001).

4 Experimental Materials and Method

In our experiments, we use the Hong Kong Laws Parallel Text (HKLaws) from the Linguistic Data Consortium, a sentence-aligned bilingual corpus with 6.5M words of English, and 9M characters of Chinese. We tagged the Chinese portion of the corpus using the CKIP tagger, and the English portion using Ratnaparkhi's tagger (Ratnaparkhi, 1996). Note that the English portion of HKLaws is about 10% of the size of the corpus used by Merlo and Stevenson (2001) in their original experiments, so we are restricted to a much smaller source of data.

Given the relatively small size of our corpus, and its narrow domain, we were only able to find a sample of 16 change-of-state and 16 creation/transformation verbs in English of sufficient frequency; see the appendix for the list of verbs used.[1] The English features for these 32 verbs were automatically extracted using regular expressions over the tagged English portion of the corpus.

The Chinese features were calculated as follows. For each English verb, we manually determined the Chinese translation in each aligned sentence to yield a collection of all (aligned) translations of the verb. This is the "aligned translation set." We also extracted all occurrences of the Chinese verbs in the aligned translation set *across the corpus*, yielding the "unaligned translation set"—i.e., the possible Chinese translations of an English target verb even when they did not occur as the translation of that verb.

The required counts for the Chinese features were collected for these verbs partly automatically (*Chinese Verb POS tags*, *Passive Particles*, *Periphrastic Particles*, and *Morpheme Length*) and partly by hand (*Semantic Specificity* and *Morpheme POS combinations*). The value of a Chinese feature for a given verb is the normalized frequency of occurrence of the feature across all occurrences of that verb in the given translation set. The resulting frequencies for the aligned translation set form the aligned dataset, and those for the unaligned translation set form the unaligned dataset.

The motivation for collecting unaligned data is to examine an alternative method for combining multilingual data. Note that parallel corpora, especially those that are sentence-aligned, are difficult to construct. Most parallel corpora we found are considerably smaller than some of the more popular monolingual ones. Given that more monolingual corpora are available, we want to explore the possibility of using non-parallel texts from multiple languages (hence, necessarily unaligned data), rather than solely looking at bilingual corpora.

[1] In the set of creation/transformation verbs, we include one item not from that class, but with similar syntactic behavior, the verb *pack*. We included this verb because we could not find another creation/transformation verb in the HKLaws corpus. We could have used another optionally intransitive (noncausative) class from Levin's classification, but wanted to focus on these two classes in order to provide maximum comparability to the ongoing work by Stevenson and Merlo, who are currently investigating these classes.

In order to compare our results to the monolingual method on a large corpus (as in (Merlo and Stevenson, 2001)), we also collected the 5 English features for our verbs from the 65M word WSJ corpus. As a result, we have a total of four data sets: English HKLaws dataset, English WSJ dataset, aligned Chinese HKLaws dataset, and unaligned Chinese HKLaws dataset. This allows us to look at four datasets individually (the two English and two Chinese sets), and to pair up the English and Chinese datasets in four different ways (each English set paired with each Chinese set).

The data for each of our machine learning experiments consists of a vector of the relevant (English and/or Chinese) features for each verb:

Template: [verb, Eng. Feats., Chi. Feats., class]
Example: [altered, 0.04, ..., 1, change-of-state]

Combining all the English and Chinese features yields a total of 22 features. We use the resulting vectors as the training data for a classifier using the same decision tree algorithm as in (Merlo and Stevenson, 2001) (C5.0, http://www.rulequest.com). We used both 8-fold cross-validation (repeated 50 times) and leave-one-out training methodologies for our experiments.[2]

For our 8-fold cross-validation experiments, we empirically tested the tuning options available in C5.0. Except for the tree pruning percentage, we found the available options offer little to no improvements over the default settings. We set the pruning factor to 30% for the best overall performance over a variety of different combinations of features. (According to the manual, the default is 25%. A larger pruning factor results in less pruning in the decision tree.)

The cross-validation experiments train on a large number of random subsets of the data, for which we report average accuracy and standard error. The goal of the cross-validation experiments is to evaluate the contribution of different features to learning, and if possible find the best feature combination(s). To do so, we varied the precise set of features used in each experiment. Since we have a total of 17 features, performing an exhaustive search of $2^{17} \approx$ 131 thousand experiments is nearly impossible. Instead, we analysed the performance of individual monolingual features alone, and their performance when combined with the features from the other language.

The leave-one-out experiments complement the cross-validation methodology: there are a small number of tests, but we have the result of classifying each verb rather than average performance data on random subsets. Our goal for the leave-one-out experiments is to compare the precision and recall across the two classes. A feature is selected for the leave-one-out experiments if it contributed highly to performance in the cross-validation experiments.

5 Experimental Results

We report here the key results of our cross-validation and leave-one-out experiments. (For additional results and details, see (Tsang, 2001).) Since our task is a two-way classification with equal-sized classes, the chance accuracy is 50%. Although the theoretical maximum accuracy is 100%, it is worth noting that, for their three-way verb classification task, (Merlo and Stevenson, 2001) experimentally determined a best performance of 87% among a group of human experts, indicating that a more realistic upper-bound for the machine-learning task falls well below 100%.

5.1 8-Fold Cross-Validation

Our cross-validation experiments fall into three general sets. In each of these types of experiments, we use various combinations of the datasets (English HKLaws, English WSJ, Chinese aligned and unaligned), as explained in detail below. First, we analysed the contribution of the English features to learning by testing all English features together, and all English features individually. These tests form our baseline results using monolingual English data. Second, we similarly analysed the contribution of the Chinese features to learning by testing all Chinese features together and all Chinese features individually. Finally, since our overall goal is to observe possible information gain by augmenting English data with non-English data, we present results in which

[2] An 8-fold cross-validation experiment divides the data into eight parts (folds) and runs eight times, each time training on a different 7/8 of the data and testing on the remaining 1/8. We chose 8 folds simply because it evenly divides our 32 verbs. In leave-one-out experiments, we leave out one vector for testing and use the remaining vectors for training, repeated 32 times (once for each verb).

Features	%Acc.	%SE
HKLaws, All English Features	41.3	0.7
HKLaws, Transitivity	49.5	0.5
WSJ, All English Features	66.3	0.6
WSJ, Animacy	72.5	0.4

Table 1: Accuracy (%Acc.) and Standard Error (%SE) in the 8-Fold Cross-Validation Experiments, Using English Features Only

Aligned Features	%Acc.	%SE	Unaligned Features	%Acc.	%SE
HKLaws, All Chi. Features	75.4	0.6	HKLaws, All Chi. Features	74.1	0.6
HKLaws, UPenn VA-Tag	75.1	0.4	HKLaws, UPenn VV-Tag	71.5	0.5

Table 2: Accuracy (%Acc.) and Standard Error (%SE) in the 8-Fold Cross-Validation Experiments, Using Chinese Features Only

we add selected Chinese features to the set of English features.

Table 1 shows the results of our experiments evaluating the English features. Using the HKLaws dataset, English features alone achieved a best performance of no better than chance (49.5% accuracy, SE 0.5%). Using the WSJ dataset, all the English features together achieved an accuracy of 66.3% (SE 0.6%), although the best performance was achieved by a single English feature alone (animacy), with an accuracy of 72.5% (SE 0.4%). We note then that the English HKLaws dataset alone is not sufficiently informative for the classification task. The best accuracy achieved with the WSJ data, of 72.5%, will serve as our monolingual baseline—i.e., the performance we would like to beat with our multilingual data.

Next, we turn to our evaluation of Chinese features alone; the results are reported in Table 2. We see that, in contrast to the English HKLaws dataset, the Chinese features alone performed very well. For the aligned and unaligned Chinese HKLaws datasets, using all Chinese features achieved an accuracy of 75.4% and 74.1%, respectively , as shown in line 1 of the table; the two results are not significantly different at the $p<0.05$ level. Using the verb POS tags alone in the aligned set—e.g., the UPenn VA (stative) tag, in line 2 of the table—achieves comparable performance of 75.1%, SE 0.4% (again, not statistically different from the first two results). (The best single feature in the unaligned dataset is also one of the verb tags, achieving only a slightly lower accuracy of 71.5%, SE 0.5%.)

Thus, we have the surprising result that Chinese features alone, from a fairly small dataset, are far superior to the English features from the same bilingual corpus (75.4% versus 49.5% best accuracy respectively). In fact, the Chinese features alone outperform the monolingual baseline of 72.5%, which uses English features from a much larger corpus. (The difference between the best English-only and best Chinese-only accuracies is small, but statistically significant at the $p<0.05$ level.)

Finally, we want to look at the performance of all English features (from either corpus) augmented with selected Chinese features (aligned or unaligned, from the HKLaws corpus). The results are shown in Table 3. In general, combining English with Chinese features performed very well. Using English HKLaws data, the best feature combination (using the Chinese CKIP POS tags) achieved a performance of 77.9% accuracy (SE 0.8%), for a reduction of 56% of the baseline error rate. (See line 1 of Table 3; the results for aligned and unaligned data are not significantly different.) Note that, although numerically larger, these results do not differ significantly from the Chinese-only results. We conclude that for the English HKLaws dataset, the Chinese features greatly help the English features, and the English features do not hurt performance of the Chinese features.

We also augmented the English WSJ dataset with the Chinese HKLaws dataset; the best accuracy is at 80.6% (SE 0.6%), for an error

Aligned Features	%Acc.	%SE	Unaligned Features	%Acc.	%SE
HKLaws only, All Eng. Features + CKIP Tags	77.5	0.7	HKLaws only, All Eng. Features + CKIP Tags	77.9	0.8
WSJ + HKLaws, All Eng. Features + UPenn VA-Tag	80.6	0.6	WSJ + HKLaws, All Eng. Features + Peri. Part.	76.2	0.6

Table 3: Accuracy (%Acc.) and Standard Error (%SE) in the 8-Fold Cross-Validation Experiments, Using a Combination of English and Chinese Features.

	Aligned			Unaligned		
	Change-of-State	Creation / Transformation	All Verbs	Change-of-State	Creation / Transformation	All Verbs
Features	F	F	%Acc. (#E)	F	F	%Acc. (#E)
Chi. Only	0.77	0.79	78.1 (7)	0.82	0.80	81.3 (6)
Eng. Only	0.63	0.63	62.5 (12)	Aligned = Unaligned		
+ 1	***0.80***	***0.82***	***81.3 (6)***	0.63	0.63	62.5 (12)
+ 2	0.58	0.61	59.4 (13)	0.73	0.76	75.0 (8)
+ 3	0.52	0.55	53.1 (15)	0.80	0.82	81.3 (6)
+ 1,2	***0.79***	***0.83***	***81.3 (6)***	***0.83***	***0.86***	***84.4 (5)***
+ 2,3	0.48	0.57	53.1 (15)	0.69	0.69	68.8 (10)
+ 1,3	***0.79***	***0.83***	***81.3 (6)***	0.57	0.67	62.5 (12)
+ 1,2,3	***0.79***	***0.83***	***81.3 (6)***	0.62	0.74	68.8 (10)

Table 4: F-measure (F), Accuracy (%Acc.), and Number of Errors (#E) in the Leave-one-out Experiments. (1 = CKIP Tags; 2 = Passive Particles; 3 = Periphrastic Particles)

rate reduction of 61% (see line 2 of Table 3). This best performance is achieved using the UPenn VA tag in the aligned corpus, shown above to be highly useful on its own. Here, the performance of the combined dataset—using both English and Chinese features—is significantly better than both the English monolingual baseline (of 72.5%), and the Chinese features alone (best accuracy of 75.4%) ($p<0.05$).

We conclude that combining multilingual data has a significant performance benefit over monolingual data from either language. In particular, in augmenting English-only data with Chinese data, we achieve higher accuracies than that using either the English HKLaws subcorpus or the much larger WSJ corpus alone. On the other hand, we found that Chinese features alone achieve very good accuracies, close to the performance of the combined datasets, indicating that the Chinese features are highly informative in and of themselves.

Finally, we note that, although the English features from the smaller bilingual corpus were not useful in classification on their own, the combination of English and Chinese features from that corpus performed comparably to the combination of English WSJ features with the Chinese features. Thus, a smaller bilingual corpus may be effectively used either alone or in combination with a larger monolingual corpus.

5.2 Leave-One-Out Methodology

For the leave-one-out experiments, we only report results using English WSJ data in conjunction with the Chinese HKLaws data, since that yielded the best performance. We focus here on augmenting the English dataset with Chinese features that seem particularly promising. Recall that since the leave-one-out method yields the result of classifying each individual verb, we can further analyse the performance within and across the two classes with this multilingual data.

For these tests, we selected the three Chinese features *CKIP Tags*, *Passive Particles*, and *Periphrastic Particles*, because they consistently had an above-chance performance, and/or improved performance when combined with other features, in the cross-validation experiments. The results are shown in Table 4. The italicized sections highlight the feature sets with the best overall accuracies. On the left panel, showing the results with aligned Chinese data, the addition to the English features of any feature combination that includes *CKIP Tags* has the (same) best overall accuracy. On the right panel, showing the unaligned data, the addition of *CKIP Tags* and *Passive Particles* has the best overall performance. We see again that with the right feature combination, using multilingual data is superior to using English-only data.

Since we know the number of errors per class, we were able to calculate the precision and recall of each of the two classes as well. Due to space limitations, we only report the F-measure in Table 4. For each class, we calculated a balanced F score as 2PR/(P+R), where P and R are the precision and recall. The two classes yield similar F scores in almost all cases, and the trend is not different from that of the overall accuracy. Observe in the italicized sections in the table (the best overall performance), the F scores are larger than those in the monolingual section (first two lines of the table). We conclude that adding Chinese features to English features has a performance benefit over the monolingual features alone for both verb classes, as well as overall.

6 Related Work

Our work is the first use of a bilingual corpus-based technique for the automatic learning of verb classification, though we are not the first to utilized multilingual resources for lexical acquisition tasks generally. For example, (Siegel and McKeown, 2000) suggested the use of parallel corpora in learning the aspectual classification (i.e., state or event) of English verbs. (Ide, 2000) and (Resnik and Yarowsky, 2000) made use of parallel corpora for word sense disambiguation. That is, a parallel (English-non-English) corpus was used as a source for lexicalizing some fine-grained English senses.

Other work using multilingual resources that is highly related to ours are studies by Fung (1998) and by Melamed et al. (1997; 1998), in which a bilingual corpus was used to extract bilingual lexical entries. An important assumption is that the bilingual corpus is sentence or segment alignable, which allows for the calculation some co-occurence score between any two possible translations. One common theme in these papers is that, given any arbitrary tokens and some text coordinate system, the closer the two tokens' coordinates are, the more likely they are translational equivalents. Although we did not use an automatic method to find translations of verbs, our aligned data collection technique is similar in spirit. We also make one further implication that is absent in these papers: in one subcorpus of a bitext, the distribution of the different senses and usages of a word should be reflected/correlated in the distribution of its translations in the other subcorpus. We have suggested that some Chinese features are related to some English features; therefore, these Chinese features should also make a similar n-way distinction between the English verb classes.

7 Conclusions

We conclude that the use of multilingual corpora, either alone or in combination with monolingual data, can be an effective aid in verb classification. The Chinese features that worked best were the (active/stative) POS tags, and the passive and causative particles—easily extractable features indicating properties that are difficult to detect in English using only simple syntactic counts. This supports our hypothesis that a second language that provides surface-level features complementing the available English features can extend the possible feature set for verb classification, allowing the use of smaller parallel corpora in place of, or in addition to, larger monolingual data sets.

We have presented some preliminary results demonstrating the benefit of using multilingual data. However, we conducted our experiments only on a small test set of 32 verbs in one language pair. To test the generality of our hypothesis, we plan to duplicate our experiments using a larger test set, and expand our investigation to other language pairs. In fact, given our success with even unaligned data, we conjecture that our approach may be greatly enhanced by using multiple monolingual corpora from different languages which differentially express semantic features relevant to verb classification.

Acknowledgements

We gratefully acknowledge the financial support of the US National Science Foundation, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto. We thank Paola Merlo for helpful discussions on the work.

Appendix

Change-of-state verbs: *alter*, *change*, *clear*, *close*, *compress*, *contract*, *cool*, *decrease*, *diminish*, *dissolve*, *divide*, *drain*, *flood*, *multiply*, *open*, *reproduce*.

Creation and transformation verbs: *build*, *clean*, *compose*, *direct*, *hammer*, *knit*, *organise*, *pack*, *paint*, *perform*, *play*, *produce*, *recite*, *stitch*, *type*, *wash*.

References

Rod Ellis. 1997. *Second Language Acquisition*. Oxford University Press, Oxford.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In *Lecture Notes in Artificial Intelligence*, pages 1–17. Springer Publisher.

Rena Helms-Park. 1997. *Building an L2 Lexicon: The Acquisition of Verb Classes Relevant to Causativization in English by Speakers of Hindi-Urdu and Vietnamese*. Ph.D. thesis, University of Toronto, Toronto, Canada.

Nancy Ide. 2000. Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34:223–234.

Shunji Inagaki. 1997. Japanese and Chinese learners' acquisition of the narrow-range rules for the dative alternation in English. *Language Learning*, 47(4):637–669.

Alan Juffs. 2000. An overview of the second language acquisition of links between verb semantics and morpho-syntax. In John Archibald, editor, *Second Language Acquisition and Linguistic Theory*, pages 170–179. Blackwell Publishers.

Maria Lapata and Chris Brew. 1999. Using subcategorization to resolve verb class ambiguity. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 266–274, College Park, MD.

Beth Levin. 1993. *English Verb Classes and Alternations : A Preliminary Investigation*. University of Chicago, Chicago.

Diana McCarthy and Anna-Leena Korhonen. 1998. Detecting verbal participation in diathesis alternations. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 1493–1495, Montreal, Canada.

I. Dan Melamed and Mitchell P. Marcus. 1998. Automatic construction of Chinese-English translation lexicons. Technical Report 98-28, University of Pennsylvania, Philadelphia, PA.

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics*, Madrid, Spain.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*. To appear.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of The Empirical Methods in Natural Language Processing Conference*, Philadelphia, PA.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

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

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Journal of Computational Linguistics*, 26(4):595–628, December.

Vivian Tsang. 2001. Second language information transfer in automatic verb classification — a preliminary investigation (to be completed). Master's thesis, University of Toronto, Toronto, Canada.