The Ups and Downs of Preposition Error Detection in ESL Writing

Joel R. Tetreault
Educational Testing Service
660 Rosedale Road
Princeton, NJ, USA
JTetreault@ets.org

Martin Chodorow
Hunter College of CUNY
695 Park Avenue
New York, NY, USA
martin.chodorow@hunter.cuny.edu

Abstract

In this paper we describe a methodology for detecting preposition errors in the writing of non-native English speakers. Our system performs at 84% precision and close to 19% recall on a large set of student essays. In addition, we address the problem of annotation and evaluation in this domain by showing how current approaches of using only one rater can skew system evaluation. We present a sampling approach to circumvent some of the issues that complicate evaluation of error detection systems.

1 Introduction

The long-term goal of our work is to develop a system which detects errors in grammar and usage so that appropriate feedback can be given to non-native English writers, a large and growing segment of the world's population. Estimates are that in China alone as many as 300 million people are currently studying English as a second language (ESL). Usage errors involving prepositions are among the most common types seen in the writing of non-native English speakers. For example, (Izumi et al., 2003) reported error rates for English prepositions that were as high as 10% in a Japanese learner corpus. Errors can involve incorrect selection ("we arrived *to* the station"), extraneous use ("he went *to* outside"), and omission ("we are fond *null* beer"). What is responsible for making preposition usage so difficult for non-native speakers?

At least part of the difficulty seems to be due to the great variety of linguistic functions that prepositions serve. When a preposition marks the argument of a predicate, such as a verb, an adjective, or a noun, preposition selection is constrained by the argument role that it marks, the noun which fills that role, and the particular predicate. Many English verbs also display alternations (Levin, 1993) in which an argument is sometimes marked by a preposition and sometimes not (e.g., "They loaded the wagon with hay" / "They loaded hay on the wagon"). When prepositions introduce adjuncts, such as those of time or manner, selection is constrained by the object of the preposition ("at length", "in time", "with haste"). Finally, the selection of a preposition for a given context also depends upon the intended meaning of the writer ("we sat at the beach", "on the beach", "near the beach", "by the beach").

With so many sources of variation in English preposition usage, we wondered if the task of selecting a preposition for a given context might prove challenging even for native speakers. To investigate this possibility, we randomly selected 200 sentences from Microsoft's Encarta Encyclopedia, and, in each sentence, we replaced a randomly selected preposition with a blank line. We then asked two native English speakers to perform a cloze task by filling in the blank with the best preposition, given the context provided by the rest of the sentence. Our results showed only about 75% agreement between the two raters, and between each of our raters and Encarta.

The presence of so much variability in preposition function and usage makes the task of the learner a daunting one. It also poses special challenges for developing and evaluating an NLP error detection system. This paper addresses both the

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 865–872
Manchester, August 2008

development and evaluation of such a system.

First, we describe a machine learning system that detects preposition errors in essays of ESL writers. To date there have been relatively few attempts to address preposition error detection, though the sister task of detecting determiner errors has been the focus of more research. Our system performs comparably with other leading systems. We extend our previous work (Chodorow et al., 2007) by experimenting with combination features, as well as features derived from the Google N-Gram corpus and Comlex (Grishman et al., 1994).

Second, we discuss drawbacks in current methods of annotating ESL data and evaluating error detection systems, which are not limited to preposition errors. While the need for annotation by multiple raters has been well established in NLP tasks (Carletta, 1996), most previous work in error detection has surprisingly relied on only one rater to either create an annotated corpus of learner errors, or to check the system's output. Some grammatical errors, such as number disagreement between subject and verb, no doubt show very high reliability, but others, such as usage errors involving prepositions or determiners are likely to be much less reliable. Our results show that relying on one rater for system evaluation can be problematic, and we provide a sampling approach which can facilitate using multiple raters for this task.

In the next section, we describe a system that automatically detects errors involving incorrect preposition selection ("We arrived *to* the station") and extraneous preposition usage ("He went *to* outside"). In sections 3 and 4, we discuss the problem of relying on only one rater for exhaustive annotation and show how multiple raters can be used more efficiently with a sampling approach. Finally, in section 5 we present an analysis of common preposition errors that non-native speakers make.

2 System

2.1 Model

We have used a Maximum Entropy (ME) classifier (Ratnaparkhi, 1998) to build a model of correct preposition usage for 34 common English prepositions. The classifier was trained on 7 million preposition contexts extracted from parts of the MetaMetrics Lexile corpus that contain textbooks and other materials for high school students. Each context was represented by 25 features consisting of the words and part-of-speech (POS) tags found in a local window of +/- two positions around the preposition, plus the head verb of the preceding verb phrase (PV), the head noun of the preceding noun phrase (PN), and the head noun of the following noun phrase (FH), among others. In analyzing the contexts, we used only tagging and heuristic phrase-chunking, rather than parsing, so as to avoid problems that a parser might encounter with ill-formed non-native text [1]. In test mode, the classifier was given the context in which a preposition occurred, and it returned a probability for each of the 34 prepositions.

2.2 Other Components

While the ME classifier constitutes the core of the system, it is only one of several processing components that refines or blocks the system's output. Since the goal of an error detection system is to provide diagnostic feedback to a student, typically a system's output is heavily constrained so that it minimizes false positives (i.e., the system tries to avoid saying a writer's preposition is used incorrectly when it is actually right), and thus does not mislead the writer.

Pre-Processing Filter: A pre-processing program skips over preposition contexts that contain spelling errors. Classifier performance is poor in such cases because the classifier was trained on well-edited text, i.e., without misspelled words. In the context of a diagnostic feedback and assessment tool for writers, a spell checker would first highlight the spelling errors and ask the writer to correct them before the system analyzed the prepositions.

Post-Processing Filter: After the ME classifier has output a probability for each of the 34 prepositions but before the system has made its final decision, a series of rule-based post-processing filters block what would otherwise be false positives that occur in specific contexts. The first filter prevents the classifier from marking as an error a case where the classifier's most probable preposition is an antonym of what the writer wrote, such as "with/without" and "from/to". In these cases, resolution is dependent on the intent of the writer and thus is outside the scope of information cap-

[1] For an example of a common ungrammatical sentence from our corpus, consider: "In consion, for some reasons, museums, particuraly known travel place, get on many people."

tured by the current feature set. Another problem for the classifier involves differentiating between certain adjuncts and arguments. For example, in the sentence "They described a part *for a kid*", the system's top choices were *of* and *to*. The benefactive adjunct introduced by *for* is difficult for the classifier to learn, perhaps because it so freely occurs in many locations within a sentence. A post-processing filter prevents the system from marking as an error a prepositional phrase that begins with *for* and has an object headed by a human noun (a WordNet hyponym of *person* or *group*).

Extraneous Use Filter: To cover extraneous use errors, we developed two rule-based filters: 1) Plural Quantifier Constructions, to handle cases such as "some *of* people" and 2) Repeated Prepositions, where the writer accidentally repeated the same preposition two or more times, such as "can find friends *with with*". We found that extraneous use errors usually constituted up to 18% of all preposition errors, and our extraneous use filters handle a quarter of that 18%.

Thresholding: The final step for the preposition error detection system is a set of thresholds that allows the system to skip cases that are likely to result in false positives. One of these is where the top-ranked preposition and the writer's preposition differ by less than a pre-specified amount. This was also meant to avoid flagging cases where the system's preposition has a score only slightly higher than the writer's preposition score, such as: "My sister usually gets home around 3:00" (writer: around = 0.49, system: by = 0.51). In these cases, the system's and the writer's prepositions both fit the context, and it would be inappropriate to claim the writer's preposition was used incorrectly. Another system threshold requires that the probability of the writer's preposition be lower than a pre-specified value in order for it to be flagged as an error. The thresholds were set so as to strongly favor precision over recall due to the high number of false positives that may arise if there is no thresholding. This is a tactic also used for determiner selection in (Nagata et al., 2006) and (Han et al., 2006). Both thresholds were empirically set on a development corpus.

2.3 Combination Features

ME is an attractive choice of machine learning algorithm for a problem as complex as preposition error detection, in no small part because of the availability of ME implementations that can handle many millions of training events and features. However, one disadvantage of ME is that it does not automatically model the interactions among features as some other approaches do, such as support vector machines (Jurafsky and Martin, 2008). To overcome this, we have experimented with augmenting our original feature set with "combination features" which represent richer contextual structure in the form of syntactic patterns.

Table 1 (first column) illustrates the four combination features used for the example context "take our place in the line". The *p* denotes a preposition, so N-*p*-N denotes a syntactic context where the preposition is preceded and followed by a noun phrase. We use the preceding noun phrase (PN) and following head (FH) from the original feature set for the N-*p*-N feature. Column 3 shows one instantiation of combination features: Combo:word. For the N-*p*-N feature, the corresponding Combo:word instantiation is "place-line" since "place" is the PN and "line" is the FH. We also experimented with using combinations of POS tags (Combo:tag) and word+tag combinations (Combo:word+tag). So for the example, the Combo:tag N-*p*-N feature would be "NN-NN", and the Combo:word+tag N-*p*-N feature would be place_NN+line_NN (see the fourth column of Table 1). The intuition with the Combo:tag features is that the Combo:word features have the potential to be sparse, and these capture more general patterns of usage.

We also experimented with other features such as augmenting the model with verb-preposition preferences derived from Comlex (Grishman et al., 1994), and querying the Google Terabyte N-gram corpus with the same patterns used in the combination features. The Comlex-based features did not improve the model, and though the Google N-gram corpus represents much more information than our 7 million event model, its inclusion improved performance only marginally.

2.4 Evaluation

In our initial evaluation of the system we collected a corpus of 8,269 preposition contexts, error-annotated by two raters using the scheme described in Section 3 to serve as a gold standard. In this study, we focus on two of the three types of preposition errors: using the incorrect preposition and using an extraneous preposition. We compared

Class	Components	Combo:word Features	Combo:tag Features
p-N	FH	line	NN
N-*p*-N	PN-FH	place-line	NN-NN
V-*p*-N	PV-PN	take-line	VB-NN
V-N-*p*-N	PV-PN-FH	take-place-line	VB-NN-NN

Table 1: Feature Examples for *take our place in the line*

different models: the baseline model of 25 features and baseline with combination features added. The precision and recall for the top performing models are shown in Table 2. These results do not include the extraneous use filter; this filter generally increased precision by as much as 2% and recall by as much as 5%.

Evaluation Metrics In the tasks of determiner and preposition selection in well-formed, native texts (such as (Knight and Chander, 1994), (Minnen et al., 2000), (Turner and Charniak, 2007) and (Gamon et al., 2008)), the evaluation metric most commonly used is *accuracy*. In these tasks, one compares the system's output on a determiner or preposition to the gold standard of what the writer originally wrote. However, in the tasks of determiner and preposition error detection, *precision* and *recall* are better metrics to use because one is only concerned with a subset of the prepositions (or determiners), those used incorrectly, as opposed to all of them in the selection task. In essence, accuracy has the problem of distorting system performance.

Results The baseline system (described in (Chodorow et al., 2007)) performed at 79.8% precision and 11.7% recall. Next we tested the different combination models: word, tag, word+tag, and all three. Surprisingly, three of the four combination models: tag, word+tag, all, did not improve performance of the system when added to the model, but using just the +Combo:word features improved recall by 1%. We use the +Combo:word model to test our sampling approach in section 4.

As a final test, we tuned our training corpus of 7 million events by removing any contexts with unknown or misspelled words, and then retrained the model. This "purge" resulted in a removal of nearly 200,000 training events. With this new training corpus, the +Combo:tag feature showed the biggest improvement over the baseline, with an improvement in both precision (+2.3%) and recall (+2.4%) to 82.1% and 14.1% respectively (last line of Table 2. While this improvement may seem small, it is in part due to the difficulty of the problem, but also the high baseline system score that was established in our prior work (Chodorow et al., 2007).

It should be noted that with the inclusion of the extraneous use filter, performance of the +Combo:tag rose to 84% precision and close to 19% recall.

Model	Precision	Recall
Baseline	79.8%	11.7%
+Combo:word	79.8%	12.8%
+Combo:tag (with purge)	82.1%	14.1%

Table 2: Best System Results on Incorrect Selection Task

2.5 Related Work

Currently there are only a handful of approaches that tackle the problem of preposition error detection in English learner texts. (Gamon et al., 2008) used a language model and decision trees to detect preposition and determiner errors in the CLEC corpus of learner essays. Their system performs at 79% precision (which is on par with our system), however recall figures are not presented thus making comparison difficult. In addition, their evaluation differs from ours in that they also include errors of omission, and their work focuses on the top twelve most frequent prepositions, while ours has greater coverage with the top 34. (Izumi et al., 2003) and (Izumi et al., 2004) used an ME approach to classify different grammatical errors in transcripts of Japanese interviews. They do not present performance of prepositions specifically, but overall performance for the 13 error types they target reached 25% precision and 7% recall. (Eeg-Olofsson and Knuttson, 2003) created a rule-based approach to detecting preposition errors in Swedish language learners (unlike the approaches presented here, which focus on English language learners), and their system performed at 25% accuracy. (Lee and Seneff, 2006) used a language model to tackle the novel problem of preposition selection in a dialogue corpus. While their performance results are quite high, 88% precision and

78% recall, it should be noted that their evaluation was on a small corpus with a highly constrained domain, and focused on a limited number of prepositions, thus making direct comparison with our approach difficult.

Although our recall figures may seem low, especially when compared to other NLP tasks such as parsing and anaphora resolution, this is really a reflection of how difficult the task is. For example, in the problem of preposition selection in native text, a baseline using the most frequent preposition (*of*) results in precision and recall of 26%. In addition, the cloze tests presented earlier indicate that even in well-formed text, agreement between native speakers on preposition selection is only 75%. In texts written by non-native speakers, rater disagreement increases, as will be shown in the next section.

3 Experiments with Multiple Raters

While developing an error detection system for prepositions is certainly challenging, given the results from our work and others, evaluation also poses a major challenge. To date, single human annotation has typically been the gold standard for grammatical error detection, such as in the work of (Izumi et al., 2004), (Han et al., 2006), (Nagata et al., 2006), (Eeg-Olofsson and Knuttson, 2003)[2]. Another method for evaluation is verification ((Gamon et al., 2008), where a human rater checks over a system's output. The drawbacks of this approach are: 1. every time the system is changed, a rater is needed to re-check the output, and 2. it is very hard to estimate recall. What these two evaluation methods have in common is that they side-step the issue of annotator reliability.

In this section, we show how relying on only one rater can be problematic for difficult error detection tasks, and in section 4, we propose a method ("the sampling approach") for efficiently evaluating a system that does not require the amount of effort needed in the standard approach to annotation.

3.1 Annotation

To create a gold-standard corpus of error annotations for system evaluation, and also to determine whether multiple raters are better than one, we trained two native English speakers with prior NLP annotation experience to annotate preposition errors in ESL text. The training was very extensive: both raters were trained on 2000 preposition contexts and the annotation manual was iteratively refined as necessary. To summarize the procedure, the two raters were shown sentences randomly selected from student essays with each preposition highlighted in the sentence. They marked each context ($\pm$2-word window around the preposition, plus the commanding verb) for grammar and spelling errors, and then judged whether the writer used an incorrect preposition, a correct preposition, or an extraneous preposition. Finally, the raters suggested prepositions that would best fit the context, even if there were no error (some contexts can license multiple prepositions).

3.2 Reliability

Each rater judged approximately 18,000 prepositions contexts, with 18 sets of 100 contexts judged by both raters for purposes of computing kappa. Despite the rigorous training regimen, kappa ranged from 0.411 to 0.786, with an overall combined value of 0.630. Of the prepositions that Rater 1 judged to be errors, Rater 2 judged 30.2% to be acceptable. Conversely, of the prepositions Rater 2 judged to be erroneous, Rater 1 found 38.1% acceptable. The kappa of 0.630 shows the difficulty of this task and also shows how two highly trained raters can produce very different judgments. Details on our annotation and human judgment experiments can be found in (Tetreault and Chodorow, 2008).

Variability in raters' judgments translates to variability of system evaluation. For instance, in our previous work (Chodorow et al., 2007), we found that when our system's output was compared to judgments of two different raters, there was a 10% difference in precision and a 5% difference in recall. These differences are problematic when evaluating a system, as they highlight the potential to substantially over- or under-estimate performance.

4 Sampling Approach

The results from the previous section motivate the need for a more refined evaluation. They suggest that for certain error annotation tasks, such as preposition usage, it may not be appropriate to use only one rater and that if one uses multiple raters

[2](Eeg-Olofsson and Knuttson, 2003) had a small evaluation on 40 preposition contexts and it is unclear whether multiple annotators were used.

for error annotation, there is the possibility of creating an adjudicated set, or at least calculating the variability of the system's performance. However, annotation with multiple raters has its own disadvantages as it is much more expensive and time-consuming. Even using one rater to produce a sizeable evaluation corpus of preposition errors is extremely costly. For example, if we assume that 500 prepositions can be annotated in 4 hours using our annotation scheme, and that the base rate for preposition errors is 10%, then it would take at least 80 hours for a rater to find and mark 1000 errors. In this section, we propose a more efficient annotation approach to circumvent this problem.

4.1 Methodology

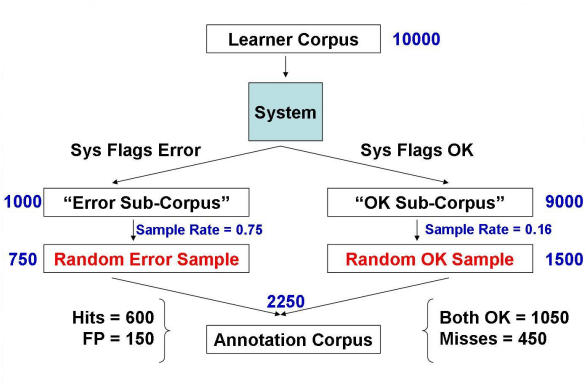

Figure 1: Sampling Approach Example

The sampling procedure outlined here is inspired by the one described in (Chodorow and Leacock, 2000) for the task of evaluating the usage of nouns, verbs and adjectives. The central idea is to skew the annotation corpus so that it contains a greater proportion of errors.

Here are the steps in the procedure:

1. Process a test corpus of sentences so that each preposition in the corpus is labeled "OK" or "Error" by the system.

2. Divide the processed corpus into two sub-corpora, one consisting of the system's "OK" prepositions and the other of the system's "Error" prepositions. For the hypothetical data in Figure 1, the "OK" sub-corpus contains 90% of the prepositions, and the "Error" sub-corpus contains the remaining 10%.

3. Randomly sample cases from each sub-corpus and combine the samples into an annotation set that is given to a "blind" human rater. We generally use a higher sampling rate for the "Error" sub-corpus because we want to "enrich" the annotation set with a larger proportion of errors than is found in the test corpus as a whole. In Figure 1, 75% of the "Error" sub-corpus is sampled while only 16% of the "OK" sub-corpus is sampled.

4. For each case that the human rater judges to be an error, check to see which sub-corpus it came from. If it came from the "OK" sub-corpus, then the case is a Miss (an error that the system failed to detect). If it came from the "Error" sub-corpus, then the case is a Hit (an error that the system detected). If the rater judges a case to be a correct usage and it came from the "Error" sub-corpus, then it is a False Positive (FP).

5. Calculate the proportions of Hits and FPs in the sample from the "Error" sub-corpus. For the hypothetical data in Figure 1, these values are 600/750 = 0.80 for Hits, and 150/750 = 0.20 for FPs. Calculate the proportion of Misses in the sample from the "OK" sub-corpus. For the hypothetical data, this is 450/1500 = 0.30 for Misses.

6. The values computed in step 5 are conditional proportions based on the sub-corpora. To calculate the overall proportions in the test corpus, it is necessary to multiply each value by the relative size of its sub-corpus. This is shown in Table 3, where the proportion of Hits in the "Error" sub-corpus (0.80) is multiplied by the relative size of the "Error" sub-corpus (0.10) to produce an overall Hit rate (0.08). Overall rates for FPs and Misses are calculated in a similar manner.

7. Using the values from step 6, calculate Precision (Hits/(Hits + FP)) and Recall (Hits/(Hits + Misses)). These are shown in the last two rows of Table 3.

	Estimated Overall Rates Sample Proportion * Sub-Corpus Proportion
Hits	0.80 * 0.10 = **0.08**
FP	0.20 * 0.10 = **0.02**
Misses	0.30 * 0.90 = **0.27**
Precision	0.08/(0.08 + 0.02) = **0.80**
Recall	0.08/(0.08 + 0.27) = **0.23**

Table 3: Sampling Calculations (Hypothetical)

This method is similar in spirit to *active learning* ((Dagan and Engelson, 1995) and (Engelson and Dagan, 1996)), which has been used to iteratively build up an annotated corpus, but it differs from active learning applications in that there are no iterative loops between the system and the human annotator(s). In addition, while our methodology is used for *evaluating* a system, active learning is commonly used for *training* a system.

4.2 Application

Next, we tested whether our proposed sampling approach provides good estimates of a system's performance. For this task, we used the +Combo:word model to separate a large corpus of student essays into the "Error" and "OK" subcorpora. The original corpus totaled over 22,000 prepositions which would normally take several weeks for two raters to double annotate and then adjudicate. After the two sub-corpora were proportionally sampled, this resulted in an annotation set of 752 preposition contexts (requiring roughly 6 hours for annotation), which is substantially more manageable than the full corpus. We had both raters work together to make judgments for each preposition.

It is important to note that while these are not the exact same essays used in the previous evaluation of 8,269 preposition contexts, they come from the same pool of student essays and were on the same topics. Given these strong similarities, we feel that one can compare scores between the two approaches. The precision and recall scores for both approaches are shown in Table 4 and are extremely similar, thus suggesting that the sampling approach can be used as an alternative to exhaustive annotation.

	Precision	Recall
Standard Approach	80%	12%
Sampling Approach	79%	14%

Table 4: Sampling Results

It is important with the sampling approach to use appropriate sample sizes when drawing from the sub-corpora, because the accuracy of the estimates of hits and misses will depend upon the proportion of errors in each sub-corpus as well as on the sample sizes. The OK sub-corpus is expected to have even fewer errors than the overall base rate, so it is especially important to have a relatively large sample from this sub-corpus. The comparison study described above used an OK sub-corpus sample that was twice as large as the Error subcorpus sample (about 500 contexts vs. 250 contexts).

In short, the sampling approach is intended to alleviate the burden on annotators when faced with the task of having to rate several thousand errors of a particular type in order to produce a sizeable error corpus. On the other hand, one advantage that exhaustive annotation has over the sampling method is that it makes possible the comparison of multiple systems. With the sampling approach, one would have to resample and annotate for each system, thus multiplying the work needed.

5 Analysis of Learner Errors

One aspect of automatic error detection that usually is under-reported is an analysis of the errors that learners typically make. The obvious benefit of this analysis is that it can focus development of the system.

From our annotated set of preposition errors, we found that the most common prepositions that learners used incorrectly were *in* (21.4%), *to* (20.8%) and *of* (16.6%). The top ten prepositions accounted for 93.8% of all preposition errors in our learner corpus.

Next, we ranked the common preposition "confusions", the common mistakes made for each preposition. The top ten most common confusions are listed in Table 5, where *null* refers to cases where no preposition is licensed (the writer used an extraneous preposition). The most common offenses were actually extraneous errors (see Table 5): using *to* and *of* when no preposition was licensed accounted for 16.8% of all errors.

It is interesting to note that the most common usage errors by learners overwhelmingly involved the ten most frequently occurring prepositions in native text. This suggests that our effort to handle the 34 most frequently occurring prepositions may be overextended and that a system that is specifically trained and refined on the top ten prepositions may provide better diagnostic feedback to a learner.

6 Conclusions

This paper has two contributions to the field of error detection in non-native writing. First, we discussed a system that detects preposition errors with high precison (up to 84%) and is competitive

Writer's Prep.	Rater's Prep.	Frequency
to	*null*	9.5%
of	*null*	7.3%
in	at	7.1%
to	for	4.6%
in	*null*	3.2%
of	for	3.1%
in	on	3.1%
of	in	2.9%
at	in	2.7%
for	to	2.5%

Table 5: Common Preposition Confusions

with other leading methods. We used an ME approach augmented with combination features and a series of thresholds. This system is currently incorporated in the *Criterion* writing evaluation service. Second, we showed that the standard approach to evaluating NLP error detection systems (comparing a system's output with a gold-standard annotation) can greatly skew system results when the annotation is done by only one rater. However, one reason why a single rater is commonly used is that building a corpus of learner errors can be extremely costly and time consuming. To address this efficiency issue, we presented a sampling approach that produces results comparable to exhaustive annotation. This makes using multiple raters possible since less time is required to assess the system's performance. While the work presented here has focused on prepositions, the arguments against using only one rater, and for using a sampling approach generalize to other error types, such as determiners and collocations.

Acknowledgements We would first like to thank our two annotators Sarah Ohls and Waverly VanWinkle for their hours of hard work. We would also like to acknowledge the three anonymous reviewers and Derrick Higgins for their helpful comments and feedback.

References

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, pages 249–254.

Chodorow, M. and C. Leacock. 2000. An unsupervised method for detecting grammatical errors. In *NAACL*.

Chodorow, M., J. Tetreault, and N-R. Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*.

Dagan, I. and S. Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Proceedings of ICML*, pages 150–157.

Eeg-Olofsson, J. and O. Knuttson. 2003. Automatic grammar checking for second language learners - the use of prepositions. In *Nodalida*.

Engelson, S. and I. Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *Proceedings of ACL*, pages 319–326.

Gamon, M., J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. 2008. Using contextual speller techniques and language modeling for esl error correction. In *IJCNLP*.

Grishman, R., C. Macleod, and A. Meyers. 1994. Comlex syntax: Building a computational lexicon. In *COLING*.

Han, N-R., M. Chodorow, and C. Leacock. 2006. Detecting errors in english article usage by non-native speakers. *Natural Language Engineering*, 12:115–129.

Izumi, E., K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara. 2003. Automatic error detection in the Japanese leaners' English spoken data. In *ACL*.

Izumi, E., K. Uchimoto, and H. Isahara. 2004. The overview of the sst speech corpus of Japanese learner English and evaluation through the experiment on automatic detection of learners' errors. In *LREC*.

Jurafsky, D. and J. Martin. 2008. *Speech and Language Processing (2nd Edition)*. Prentice Hall. To Appear.

Knight, K. and I. Chander. 1994. Automated postediting of documents. In *Conference on Artificial Intelligence*.

Lee, J. and S. Seneff. 2006. Automatic grammar correction for second-language learners. In *Interspeech*.

Levin, B. 1993. *English verb classes and alternations: a preliminary investigation*. Univ. of Chicago Press.

Minnen, G., F. Bond, and A. Copestake. 2000. Memory-based learning for article generation. In *CoNLL*.

Nagata, R., A. Kawai, K. Morihiro, and N. Isu. 2006. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proceedings of the ACL/COLING*.

Ratnaparkhi, A. 1998. *Maximum Entropy Models for natural language ambiguity resolution*. Ph.D. thesis, University of Pennsylvania.

Tetreault, J. and M. Chodorow. 2008. Native Judgments of non-native usage: Experiments in preposition error detection. In *COLING Workshop on Human Judgments in Computational Linguistics*.

Turner, J. and E. Charniak. 2007. Language modeling for determiner selection. In *HLT/NAACL*.