Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection

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

Evaluation and annotation are two of the greatest challenges in developing NLP instructional or diagnostic tools to mark grammar and usage errors in the writing of non-native speakers. Past approaches have commonly used only one rater to annotate a corpus of learner errors to compare to system output. In this paper, we show how using only one rater can skew system evaluation and then we present a sampling approach that makes it possible to evaluate a system more efficiently.

1 Introduction

In this paper, we present a series of experiments that explore the reliability of human judgments in rating preposition usage. While one tends to think of annotator disagreements about discourse and semantics as being quite common, our studies show that judgments of preposition usage, which is largely lexically driven, can be just as contentious. As a result, this unreliability poses a serious issue for the development and evaluation of NLP tools in the task of automatically detecting preposition usage errors in the writing of non-native speakers of English.

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), (Gamon et al., 2008)[1]. Although there are several learner corpora annotated for preposition and determiner errors (such as the Cambridge Learners Corpus[2] and the Chinese Learner English Corpus[3]), it is unclear which portions of these, if any, were doubly annotated. This previous work has side-stepped the issue of annotator reliability, which we address here through the following three contributions:

- **Judgments of Native Usage** To motivate our work in non-native usage, we first illustrate the difficulty of preposition selection with two experiments: a cloze test and a choice test, where native speakers judge native texts (section 4).
- **Judgments of Non-Native Usage** As stated earlier, most computational work in the field of error detection tools for non-native speakers has relied on a single rater to annotate a gold standard corpus to check a system's output. We conduct an extensive double-annotation evaluation to measure inter-rater reliability and show that using one rater can be unreliable and may produce misleading results in a system test (section 5).
- **Sampling Approach** Multiple annotation can be very costly and time-consuming, which may explain why previous work employed only one rater. As an alternative to the standard exhaustive annotation, we propose a sampling approach in which estimates of the rates of hits, false positives, and misses are derived from random samples of the system's output, and then precision and recall of the system can be calculated. We show that estimates of system performance derived

[1] (Eeg-Olofsson and Knuttson, 2003) had a small evaluation of 40 prepositions and it is unclear whether they used multiple annotators or not.

[2] http://www.cambridge.org/elt

[3] http://langbank.engl.polyu.edu.hk/corpus/clec.html

from the sampling approach are comparable to those derived from an exhaustive annotation, but require only a fraction of the effort (section 6).

In short, through a battery of experiments we show how rating preposition usage, in either native or non-native texts, is a task that has surprisingly low inter-annotator reliability and thus greatly impacts system evaluation. We then describe a method for efficiently annotating non-native texts to make multiple annotation more feasible.

In section 2, we discuss in more depth the motivation for detecting usage errors in non-native writing, as well as the complexities of preposition usage. In section 3, we describe a system that automatically detects preposition errors involving incorrect selection and extraneous usage. In sections 4 and 5 respectively, we discuss experiments on the reliability of judging native and non-native preposition usage. In section 6, we present results of our system and results from comparing the sampling approach with the standard approach of exhaustive annotation.

2 Motivation

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 foreign language. Even in predominantly English-speaking countries, the proportion of non-native speakers can be very substantial. For example, the US National Center for Educational Statistics (2002) reported that nearly 10% of the students in the US public school population speak a language other than English and have limited English proficiency . At the university level in the US, there are estimated to be more than half a million foreign students whose native language is not English (Burghardt, 2002). Clearly, there is an increasing demand for tools for instruction in English as a Second Language (ESL).

Some of the most common types of ESL usage errors involve prepositions, determiners and collocations. In the work discussed here, we target preposition usage errors, specifically those of incorrect selection ("we arrived *to* the station") and extraneous use ("he went *to* outside")[4]. Preposition errors account for a substantial proportion of all ESL usage errors. For example, (Bitchener et al., 2005) found that preposition errors accounted for 29% of all the errors made by intermediate to advanced ESL students. In addition, such errors are relatively common. In our learner corpora, we found that 6% of all prepositions were incorrectly used. Some other estimates are even higher: for example, (Izumi et al., 2003) reported error rates that were as high as 10% in a Japanese learner corpus.

At least part of the difficulty in mastering prepositions seems to be due to the great variety of linguistic functions that they 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 intention of the writer ("we sat at the beach", "on the beach", "near the beach", "by the beach").

3 Automatically Detecting Preposition Usage Errors

In this section, we give a description of our system and compare its performance to other systems. Although the focus of this paper is on human judgments in the task of error detection, we describe our system to show that variability in human judgments can impact the evaluation of a system in this task. A full description of our system and its performance can be found in (Tetreault and Chodorow, 2008).

3.1 System

Our approach treats preposition error detection as a classification problem: that is, given a context of two words before and two words after the writer's preposition, what is the best preposition to use?

[4]There is a third error type, omission ("we are fond *null* beer"), that is a topic for our future research.

An error is marked when the system's suggestion differs from the writer's by a certain threshold amount.

We have used a maximum entropy (ME) classifier (Ratnaparkhi, 1998) to select the most probable preposition for a given context from a set of 34 common English prepositions. One advantage of using ME is that there are implementations of it which can handle very large models built from millions of training events and consisting of hundreds of thousands of feature-value pairs. To construct a model, we begin with a training corpus that is POS-tagged and heuristically chunked into noun phrases and verb phrases[5]. For each preposition that occurs in the training corpus, a preprocessing program extracts a total of 25 features. These consist of words and POS tags in positions adjacent to the preposition and in the heads of nearby phrases. In addition, we include combination features that merge the head features. We also include features representing only the tags to be able to cover cases in testing where the words in the context were not seen in training.

In many NLP tasks (parsing, POS-tagging, pronoun resolution), it is easy to acquire training data that is similar to the testing data. However, in the case of grammatical error detection, one does not have that luxury because reliable error-annotated ESL corpora that are large enough for training a statistical classifier simply do not exist. To circumvent this problem, we have trained our classifier on examples of prepositions used correctly, as in news text.

3.2 Evaluation

Before evaluating our system on non-native writing, we evaluated how well it does on the task of preposition selection in native text, an area where there has been relatively little work to date. In this task, the system predicts the writer's preposition based on its context. Its prediction is scored automatically by comparison to what the writer actually wrote. Most recently, (Gamon et al., 2008) addressed preposition selection by developing a system that combined a decision tree and a language model. Besides the difference in algorithms, there is also a difference in coverage between their system, which selects among 13 prepositions plus a category for *Other*, and the system presented here, which selects among 34 prepositions. In their system evaluation, they split a corpus of Reuters News text and Microsoft Encarta into two sets: 70% for training (3.2M examples), and the remaining 30% for testing (1.4M examples). For purposes of comparison, we used the same corpus and evaluation method. While (Gamon et al., 2008) do not present their overall accuracy figures on the Encarta evaluation, they do present the precision and recall scores for each preposition. In Table 3.2, we display their results in terms of F-measures and show the performance of our system for each preposition. Our model outperforms theirs for 9 out of the 10 prepositions that both systems handle. Overall accuracy for our system is 77.4% and increases to 79.0% when 7M more training examples are added. For comparison purposes, using a majority baseline (always selecting the preposition *of*) in this domain results in an accuracy of 27.2%.

Prep	(Gamon et al., 2008)	(Tetreault et al., 2008)
in	0.592	**0.845**
for	0.459	**0.698**
of	0.759	**0.906**
on	0.322	**0.751**
to	0.627	**0.775**
with	0.361	**0.675**
at	0.372	**0.685**
by	0.502	**0.747**
as	0.699	**0.711**
from	0.528	**0.591**
about	**0.800**	0.654

Table 1: Comparison of F-measures on Encarta/Reuters Corpus

(Felice and Pullman, 2007) used perceptron classifiers for preposition selection in BNC News Text at 85% accuracy. For each of the five most frequent prepositions, they used a separate binary classifier to decide whether that preposition should be used or not. The classifiers are not combined into a unified model. When we reconfigured our system and evaluation to be comparable to (Felice and Pullman, 2007), our model achieved an accuracy of 90% on the same five prepositions when tested on Wall Street Journal News, which is similar, though not identical, to BNC News.

While systems can perform at close to 80% accuracy in the task of preposition selection in native texts, this high performance does not transfer to the end-task of detecting preposition errors in essays by non-native writers. For example, (Izumi et al., 2003) reported precision and recall as low as 25% and 7% respectively when detecting different

[5]We have avoided parsing because our ultimate test corpus is non-native writing, text that is difficult to parse due to the presence of numerous errors in spelling and syntax.

grammar errors (one of which was prepositions) in English essays by non-native writers. (Gamon et al., 2008) reported precision up to 80% in their evaluation on the CLEC corpus, but no recall figure was reported. We have found that our system (the model which performs at 77.4%), also performs as high as 80% precision, but recall ranged from 12% to 26% depending on the non-native test corpus.

While 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. In addition, in error detection tasks, high precision (and thus low recall) is favored since one wants to minimize the number of false positives a student may see. This is a common practice in grammatical error detection applications, such as in (Han et al., 2006) and (Gamon et al., 2008).

4 Human Judgments of Native Usage

4.1 Cloze Test

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. 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. In addition, we had our system predict which preposition should fill each blank as well. Our results (Table 2) showed only about 76% agreement between the two raters (bottom row), and between 74% and 78% when each rater was compared individually with the original preposition used in Encarta. Surprisingly, the system performed just as well as the two native raters, when compared with Encarta (third row). Although these results seem very promising, it should be noted that in many cases where the system disagreed with Encarta, its prediction was not a good fit for the context. But in the cases where the raters disagreed with Encarta, their prepositions were also licensed by the context, and thus were acceptable alternatives to the preposition that was used in the text.

Our cloze study shows that even with well-formed text, native raters can disagree with each other by 25% in the task of preposition selection. We can expect even more disagreement when the task is preposition error detection in "noisy" learner texts.

	Agreement	Kappa
Encarta vs. Rater 1	0.78	0.73
Encarta vs. Rater 2	0.74	0.68
Encarta vs. System	0.75	0.68
Rater 1 vs. Rater 2	0.76	0.70

Table 2: Cloze Experiment on Encarta

4.2 Choice Test

The cloze test presented above was scored by automatically comparing the system's choice (or the rater's choice) with the preposition that was actually written. But there are many contexts that license multiple prepositions, and in these cases, requiring an exact match is too stringent a scoring criterion.

To investigate how the exact match metric might underestimate system performance, and to further test the reliability of human judgments in native text, we conducted a choice test in which two native English speakers were presented with 200 sentences from Encarta and were asked to select which of two prepositions better fit the context. One was the originally written preposition and the other was the system's suggestion, displayed in random order. The human raters were also given the option of marking both prepositions as equally good or equally bad. The results indicated that both Rater 1 and Rater 2 considered the system's preposition equal to or better than the writer's preposition in 28% of the cases. This suggests that 28% of the mismatched cases in the automatic evaluation are not system errors but rather are instances where the context licenses multiple prepositions. If these mismatches in the automatic evaluation are actually cases of correct system performance, then the Encarta/Reuters test which performs at 75% accuracy (third row of Table 2), is more realistically around 82% accuracy (28% of the 25% mismatch rate is 7%).

5 Annotator Reliability

In this section, we address the central problem of evaluating NLP error detection tools on learner data. As stated earlier, most previous work has relied on only one rater to either create an annotated

corpus of learner errors, or to check the system's output. While some grammatical errors, such as number disagreement between subject and verb, no doubt show very high reliability, others, such as usage errors involving prepositions or determiners are likely to be much less reliable. In section 5.1, we describe our efforts in annotating a large corpus of student learner essays for preposition usage errors. Unlike previous work such as (Izumi et al., 2004) which required the rater to check for almost 40 different error types, we focus on annotating only preposition errors in hopes that having a single type of target will insure higher reliability by reducing the cognitive demands on the rater. Section 5.2 asks whether, under these conditions, one rater is acceptable for this task. In section 6, we describe an approach to efficiently evaluating a system that does not require the amount of effort needed in the standard approach to annotation.

5.1 Annotation Scheme

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 to annotate preposition errors in ESL text. Both annotators had prior experience in NLP annotation and also in ESL error detection. The training was very extensive: both raters were trained on 2000 preposition contexts and the annotation manual was iteratively refined as necessary. To our knowledge, this is the first scheme that specifically targets annotating preposition errors[6].

The two raters were shown sentences randomly selected from student essays, with each preposition highlighted in the sentence. The raters were also shown the sentence which preceded the one containing the preposition that they rated. The annotator was first asked to indicate if there were any spelling errors within the context of the preposition ($\pm$2-word window and the commanding verb). Next the annotator noted determiner or plural errors in the context, and then checked if there were any other grammatical errors (for example, wrong verb form). The reason for having the annotators check spelling and grammar is that other modules in a grammatical error detection system would be responsible for these error types. For an example of a sentence with multiple spelling, grammatical and collocational errors, consider the following sentence: "In consion, for some reasons, museums, particuraly known travel place, get on many people." A spelling error follows the preposition *In*, and a collocational error surrounds *on*. If the contexts are not corrected, it is impossible to discern if the prepositions are correct. Of course, there is the chance that by removing these we will screen out cases where there are multiple interacting errors in the context that involve prepositions. When comparing human judgments to the performance of the preposition module, the latter should not be penalized for other kinds of errors in the context.

Finally, the annotator judged the writer's preposition with a rating of "0-extraneous preposition", "1-incorrect preposition", "2-correct preposition", or "e-equally good prepositions". If the writer used an incorrect preposition, the rater supplied the best preposition(s) given the context. Very often, when the writer's preposition was correct, several other prepositions could also have occurred in the same context. In these cases, the annotator was instructed to use the "e" category and list the other equally plausible alternatives. After judging the use of the preposition and, if applicable, supplying alternatives, the annotator indicated her confidence in her judgment on a 2-point scale of "1-low" and "2-high".

5.2 Two Raters vs. One?

Following training, each annotator judged approximately 18,000 occurrences of preposition use. Annotation of 500 occurrences took an average of 3 to 4 hours. In order to calculate agreement and kappa values, we periodically provided identical sets of 100 preposition occurrences for both annotators to judge (totaling 1800 in all). After removing instances where there were spelling or grammar errors, and after combining categories "2" and "e", both of which were judgments of correct usage, we computed the kappa values for the remaining doubly judged sets. These ranged from 0.411 to 0.786, with an overall combined value of 0.630[7]. The confusion matrix for the combined set (totaling 1336 contexts) is shown in Table 3. The rows represent Rater 1's (R1) judgments while the columns represent Rater 2's judgments. As one

[6](Gamon et al., 2008) did not have a scheme for annotating preposition errors to create a gold standard corpus, but did use a scheme for the similar problem of verifying a system's output in preposition error detection.

[7]When including spelling and grammar annotations, kappa ranged from 0.474 to 0.773.

would expect given the prior reports of preposition error rates in non-native writing, the raters' agreement for this task was quite high overall (0.952) due primarily to the large agreement count where both annotators rated the usage "OK" (1213 total contexts). However there were 42 prepositions that both raters marked as a "Wrong Choice" and 17 as "Extraneous." It is important to note the disagreements in judging these errors: for example, Rater 1 judged 26 prepositions to be errors that Rater 2 judged to be OK, for a disagreement rate of .302 (26/86). Similarly, Rater 2 judged 37 prepositions to be errors that Rater 1 judged to be OK, for a disagreement rate of .381 (37/97).

R1↓; R2→	Extraneous	Wrong-Choice	OK
Extraneous	**17**	0	6
Wrong-Choice	1	**42**	20
OK	4	33	**1213**

Table 3: Confusion Matrix

The kappa of 0.630 and the off-diagonal cells in the confusion matrix both show the difficulty of this task and also show how two highly trained raters can produce very different judgments. This suggests that for certain error annotation tasks, such as preposition usage, it may not be appropriate to use only one rater and that using two or more raters to produce an adjudicated gold-standard set is the more acceptable path.

As a second test, we used a set of 2,000 preposition contexts from ESL essays (Chodorow et al., 2007) that were doubly annotated by native speakers with a scheme similar to that described above. We then compared an earlier version of our system to both raters' judgments, and found that there was a 10% difference in precision and a 5% difference in recall between the two system/rater comparisons. That means that if one is using only a single rater as a gold standard, there is the potential to over- or under-estimate precision by as much as 10%. Clearly this is problematic when evaluating a system's performance. The results are shown in Table 4.

	Precision	Recall
System vs. Rater 1	0.78	0.26
System vs. Rater 2	0.68	0.21

Table 4: Rater/System Comparison

6 Sampling Approach

If one uses multiple raters for error annotation, there is the possibility of creating an adjudicated set, or at least calculating the variability of system evaluation. However, annotation with multiple raters has its own disadvantages in that 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 error rate for prepositions 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.

6.1 Methodology

The sampling procedure outlined here is inspired by the one described in (Chodorow and Leacock, 2000). The central idea is to skew the annotation corpus so that it contains a greater proportion of errors. The result is that an annotator checks more potential errors since he or she is spending less time checking prepositions used correctly.

Here are the steps in the procedure. Figure 1 illustrates this procedure with a hypothetical corpus of 10,000 preposition examples.

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 subcorpora, 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 subcorpus 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.

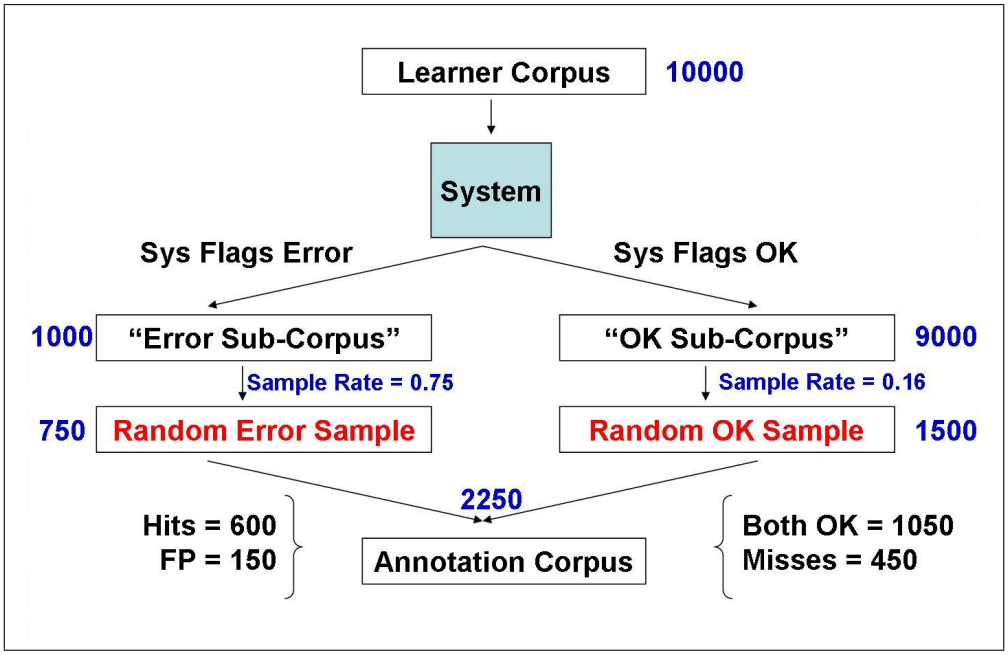

Figure 1: Sampling Approach (with hypothetical sample calculations)

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 5, 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 5.

	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 5: 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.

6.2 Application

Next, we tested whether our proposed sampling approach provides good estimates of a system's performance. For this task, we split a large corpus of ESL essays into two sets: first, a set of 8,269 preposition contexts (standard approach corpus) to be annotated using the scheme in section 5.1, and

second, a set of 22,000 preposition contexts to be rated using the sampling approach (sampling corpus). We used two non-overlapping sets because the raters were the same for this test of the two approaches.

Using the standard approach, the sampling corpus of 22,000 prepositions would normally take several weeks for two raters to double annotate and then adjudicate. After this corpus was divided into "OK" and "Error" sub-corpora, the two sub-corpora were proportionally sampled, resulting in an annotation set of 750 preposition contexts (500 contexts from the "OK" sub-corpus and 250 contexts from the "Error" sub-corpus). This required roughly 6 hours for annotation, which is substantially more manageable than the standard approach. We had both raters work together to make judgments for each preposition context.

The precision and recall scores for both approaches are shown in Table 6 and are quite similar, thus suggesting that the sampling approach can be used as an alternative to exhaustive annotation.

	Precision	Recall
Standard Approach	0.80	0.12
Sampling Approach	0.79	0.14

Table 6: Sampling Results

6.3 Confidence Intervals

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 sub-corpus sample.

One can compute the 95% confidence interval (CI) for the estimated rates of hits, misses and false positives by using the formula:

$$CI = p \pm 1.96 \times \sigma_p$$

where p is the proportion and σ_p is the standard error of the proportion given by:

$$\sigma_p = \sqrt{\frac{p(1-p)}{N}}$$

where N is the sample size.

For the example in Figure 1, the confidence interval for the proportion of Hits from the sample of the "Error" sub-corpus is:

$$CI_{hits} = 0.80 \pm 1.96 \times \sqrt{\frac{0.8 \times (1-0.80)}{750}}$$

which yields an interval of 0.077 and 0.083. Using these values, the confidence interval for precision is 0.77 to 0.83. The interval for recall can be computed in a similar manner. Of course, a larger sample size will yield narrower confidence intervals.

6.4 Summary

Table 7 summarizes the advantages and disadvantages of three methods for evaluating error detection systems. The standard (or exhaustive) approach refers to the method of annotating the errors in a large corpus. Its advantage is that the annotated corpus can be reused to evaluate the same system or compare multiple systems. However, it is costly and time-consuming which often precludes the use of multiple raters. The verification method (as used in (Gamon et al., 2008)), refers to the method of simply checking the acceptability of system output with respect to the writer's preposition. Like the sampling method, it has the advantages of efficiency and use of multiple raters (when compared to the standard method). But the disadvantage of verification is that it does not permit estimation of recall. Both verification and vampling methods require re-annotation for system retesting and comparison. In terms of system development, sampling (and to a lesser extent, verification) allows one to quickly assess system performance on a new corpus.

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 to produce a sizeable error corpus.

7 Conclusions

In this paper, we showed that the standard approach to evaluating NLP error detection systems (comparing the 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

Approach	Advantages	Disadvantages
Standard	Easy to retest system (no re-annotation required) Easy to compare systems Most reliably estimates precision and recall	Costly Time-Consuming Difficult to use multiple raters
Sampling	Efficient, especially for low-frequency errors Permits estimation of precision and recall More easily allows use of multiple raters	Less reliable estimate of recall Hard to re-test system (re-annotation required) Hard to compare systems
Verification	Efficient, especially for low-frequency errors More easily allows use of multiple raters	Does not permit estimation of recall Hard to re-test system (re-annotation required) Hard to compare systems

Table 7: Comparison of Evaluation Methods

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 reasons for using multiple raters and a sampling approach apply equally to other error types, such as determiners and collocations.

It should be noted that the work here uses two raters. For future work, we plan on annotating preposition errors with more than two raters to derive a range of judgments. We also plan to look at the effects of feedback for errors involving prepositions and determiners, on the quality of ESL writing.

The preposition error detection system described here was recently integrated into *Criterion*^{SM} Online Writing Evaluation Service developed by Educational Testing Service.

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

Bitchener, J., S. Young, and D. Cameron. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*.

Burghardt, L. 2002. Foreign applications soar at universities. *New York Times*, April.

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.

Felice, R. De and S. Pullman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*.

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*.

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*.

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

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*.

NCES. 2002. National center for educational statistics: Public school student counts, staff, and graduate counts by state: School year 2000-2001.

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

Tetreault, J. and M. Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *COLING*.