Detection of Grammatical Errors Involving Prepositions

Martin Chodorow
Hunter College of CUNY
695 Park Avenue
New York, NY, 10021
mchodoro@hunter.cuny.edu

Joel R. Tetreault and **Na-Rae Han**
Educational Testing Services
Rosedale Road
Princeton, NJ, 08541
jtetreault|nzhan@ets.org

Abstract

This paper presents ongoing work on the detection of preposition errors of non-native speakers of English. Since prepositions account for a substantial proportion of all grammatical errors by ESL (English as a Second Language) learners, developing an NLP application that can reliably detect these types of errors will provide an invaluable learning resource to ESL students. To address this problem, we use a maximum entropy classifier combined with rule-based filters to detect preposition errors in a corpus of student essays. Although our work is preliminary, we achieve a precision of 0.8 with a recall of 0.3.

1 Introduction

The National Clearinghouse for English Language Acquisition (2002) estimates that 9.6% of the students in the US public school population speak a language other than English and have limited English proficiency. Clearly, there is a substantial and increasing need for tools for instruction in English as a Second Language (ESL).

In particular, preposition usage is one of the most difficult aspects of English grammar for non-native speakers to master. Preposition errors account for a significant proportion of all ESL grammar errors. They represented the largest category, about 29%, of all the errors by 53 intermediate to advanced ESL students (Bitchener et al., 2005), and 18% of all errors reported in an intensive analysis of one Japanese writer (Murata and Ishara, 2004). Preposition errors are not only prominent among error types, they are also quite frequent in ESL writing. Dalgish (1985) analyzed the essays of 350 ESL college students representing 15 different native languages and reported that preposition errors were present in 18% of sentences in a sample of text produced by writers from first languages as diverse as Korean, Greek, and Spanish.

The goal of the research described here is to provide software for detecting common grammar and usage errors in the English writing of non-native English speakers. Our work targets errors involving prepositions, specifically those of incorrect preposition selection, such as *arrive <u>to</u> the town*, and those of extraneous prepositions, as in *most <u>of</u> people*.

We present an approach that combines machine learning with rule-based filters to detect preposition errors in a corpus of ESL essays. Even though this is work in progress, we achieve precision of 0.8 with a recall of 0.3. The paper is structured as follows: in the next section, we describe the difficulty in learning English preposition usage; in Section 3, we discuss related work; in Sections 4-7 we discuss our methodology and evaluation.

2 Problem of Preposition Usage

Why are prepositions so difficult to master? Perhaps it is because they perform so many complex roles. In English, prepositions appear in adjuncts, they mark the arguments of predicates, and they combine with other parts of speech to express new meanings.

The choice of preposition in an adjunct is largely constrained by its object (*<u>in</u> the summer*, *<u>on</u> Friday*,

Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions, pages 25–30,
Prague, Czech Republic, June 2007.

at noon) and the intended meaning (*at the beach*, *on the beach*, *near the beach*, *by the beach*). Since adjuncts are optional and tend to be flexible in their position in a sentence, the task facing the learner is quite complex.

Prepositions are also used to mark the arguments of a predicate. Usually, the predicate is expressed by a verb, but sometimes it takes the form of an adjective (*He was fond of beer*), a noun (*They have a thirst for knowledge*), or a nominalization (*The child's removal from the classroom*). The choice of the preposition as an argument marker depends on the type of argument it marks, the word that fills the argument role, the particular word used as the predicate, and whether the predicate is a nominalization. Even with these constraints, there are still variations in the ways in which arguments can be expressed. Levin (1993) catalogs verb alternations such as *They loaded hay on the wagon* vs. *They loaded the wagon with hay*, which show that, depending on the verb, an argument may sometimes be marked by a preposition and sometimes not.

English has hundreds of phrasal verbs, consisting of a verb and a particle (some of which are also prepositions). To complicate matters, phrasal verbs are often used with prepositions (i.e., *give up on someone*; *give in to their demands*). Phrasal verbs are particularly difficult for non-native speakers to master because of their non-compositionality of meaning, which forces the learner to commit them to rote memory.

3 Related Work

If mastering English prepositions is a daunting task for the second language learner, it is even more so for a computer. To our knowledge, only three other groups have attempted to automatically detect errors in preposition usage. Eeg-Olofsson et al. (2003) used 31 handcrafted matching rules to detect extraneous, omitted, and incorrect prepositions in Swedish text written by native speakers of English, Arabic, and Japanese. The rules, which were based on the kinds of errors that were found in a training set of text produced by non-native Swedish writers, targeted spelling errors involving prepositions and some particularly problematic Swedish verbs. In a test of the system, 11 of 40 preposition errors were correctly detected.

Izumi et al. (2003) and (2004) used error-annotated transcripts of Japanese speakers in an interview-based test of spoken English to train a maximum entropy classifier (Ratnaparkhi, 1998) to recognize 13 different types of grammatical and lexical errors, including errors involving prepositions. The classifier relied on lexical and syntactic features. Overall performance for the 13 error types reached 25.1% precision with 7.1% recall on an independent test set of sentences from the same source, but the researchers do not separately report the results for preposition error detection. The approach taken by Izumi and colleagues is most similar to the one we have used, which is described in the next section.

More recently, (Lee and Seneff, 2006) used a language model and stochastic grammar to replace prepositions removed from a dialogue corpus. Even though they reported a precision of 0.88 and recall of 0.78, their evaluation was on a very restricted domain with only a limited number of prepositions, nouns and verbs.

4 The Selection Model

A preposition error can be a case of incorrect preposition selection (*They arrived to the town*), use of a preposition in a context where it is prohibited (*They came to inside*), or failure to use a preposition in a context where it is obligatory (e.g., *He is fond this book*). To detect the first type of error, incorrect selection, we have employed a maximum entropy (ME) model to estimate the probability of each of 34 prepositions, based on the features in their local contexts. The ME Principle says that the best model will satisfy the constraints found in the training, and for those situations not covered in the training, the best model will assume a distribution of maximum entropy. This approach has been shown to perform well in combining heterogeneous forms of evidence, as in word sense disambiguation (Ratnaparkhi, 1998). It also has the desirable property of handling interactions among features without having to rely on the assumption of feature independence, as in a Naive Bayesian model.

Our ME model was trained on 7 million "events" consisting of an outcome (the preposition that appeared in the training text) and its associated context (the set of feature-value pairs that accompanied it). These 7 million prepositions and their contexts were extracted from the MetaMetrics corpus of 1100 and 1200 Lexile text (11th and 12th grade) and newspaper text from the San Jose Mercury News. The sentences were then POS-tagged (Ratnaparkhi, 1998) and then chunked into noun phrases and verb phrases by a heuristic chunker.

The maximum entropy model was trained with 25 contextual features. Some of the features represented the words and tags found at specific locations adjacent to the preposition; others represented the head words and tags of phrases that preceded or followed the preposition. Table 1 shows a subset of the feature list.

Some features had only a few values while others had many. PHR_pre is the "preceding phrase" feature that indicates whether the preposition was preceded by a noun phrase (NP) or a verb phrase (VP). In the example in Table 2, the preposition *into* is preceded by an NP. In a sentence that begins *After the crowd was whipped up into a frenzy of anticipation*, the preposition *into* is preceded by a VP. There were only two feature#value pairs for this feature: PHR_pre#NP and PHR_pre#VP. Other features had hundreds or even thousands of different values because they represented the occurrence of specific words that preceded or followed a preposition. Any feature#value pairs which occurred with very low frequency in the training (less than 10 times in the 7 million contexts) were eliminated to avoid the need for smoothing their probabilities. Lemma forms of words were used as feature values to further reduce the total number and to allow the model to generalize across inflectional variants. Even after incorporating these reductions, the number of values was still very large. As Table 1 indicates, TGR, the word sequence including the preposition and the two words to its right, had 54,906 different values. Summing across all features, the model contained a total of about 388,000 feature#value pairs. Table 2 shows an example of where some of the features are derived from.

5 Evaluation on Grammatical Text

The model was tested on 18,157 preposition contexts extracted from 12 files randomly selected from a portion of 1100 Lexile text (11th grade) that had not been used for training. For each context, the model predicted the probability of each preposition given the contextual representation. The highest probability preposition was then compared to the preposition that had actually been used by the writer. Because the test corpus consisted of published, edited text, we assumed that this material contained few, if any, errors. In this and subsequent tests, the model was used to classify each context as one of 34 classes (prepositions).

Results of the comparison between the classifier and the test set showed that the overall proportion of agreement between the text and the classifier was 0.69. The value of kappa was 0.64. When we examined the errors, we discovered that, frequently, the classifier's most probable preposition (the one it assigned) differed from the second most probable by just a few percentage points. This corresponded to a situation in which two or more prepositions were likely to be found in a given context. Consider the context *They thanked him for his consideration* ___ *this matter*, where either *of* or *in* could fill the blank. Because the classifier was forced to make a choice in this and other close cases, it incurred a high probability of making an error. To avoid this situation, we re-ran the test allowing the classifier to skip any preposition if its top ranked and second ranked choices differed by less than a specified amount. In other words, we permitted it to respond only when it was confident of its decision. When the difference between the first and second ranked choices was 0.60 or greater, 50% of the cases received no decision, but for the remaining half of the test cases, the proportion of agreement was 0.90 and kappa was 0.88. This suggests that a considerable improvement in performance can be achieved by using a more conservative approach based on a higher confidence level for the classifier.

6 Evaluation on ESL Essays

To evaluate the ME model's suitability for analyzing ungrammatical text, 2,000 preposition contexts were extracted from randomly selected essays written on ESL tests by native speakers of Chinese, Japanese, and Russian. This set of materials was used to look for problems that were likely to arise as a conse-

Feature	**Description**	**No. of values with freq $\geq$ 10**
BGL	Bigram to left; includes preceding word and POS	23,620
BGR	Bigram to right; includes following word and POS	20,495
FH	Headword of the following phrase	19,718
FP	Following phrase	40,778
PHR_pre	Preceding phrase type	2
PN	Preceding noun	18,329
PNMod	Adjective modifying preceding noun	3,267
PNP	Preceding noun phrase	29,334
PPrep	Preceding preposition	60
PV	Preceding verb	5,221
PVP	Preceding verb phrase	23,436
PVtag	POS tag of the preceding verb	24
PVword	Lemma of the preceding verb	5,221
PW	Lemma of the preceding word	2,437
TGL	Trigram to left; includes two preceding words and POS	44,446
TGR	Trigram to right; includes two following words and POS	54,906

Table 1: Some features used in ME Model

After	whipping	the	crowd	up	**into**	a	frenzy	of	anticipation...
	PVword		PN	PW			FH		
				BGL		BGR			
		——TGL——				——TGR——			

Table 2: Locations of some features in the local context of a preposition

quence of the mismatch between the training corpus (edited, grammatical text) and the testing corpus (ESL essays with errors of various kinds). When the model was used to classify prepositions in the ESL essays, it became obvious, almost immediately, that a number of new performance issues would have to be addressed.

The student essays contained many misspelled words. Because misspellings were not in the training, the model was unable to use the features associated with them (e.g., FHword#frinzy) in its decision making. The tagger was also affected by spelling errors, so to avoid these problems, the classifier was allowed to skip any context that contained misspelled words in positions adjacent to the preposition or in its adjacent phrasal heads. A second problem resulted from punctuation errors in the student writing. This usually took the form of missing commas, as in *I disagree because from my point of view there is no evidence*. In the training corpus, commas generally separated parenthetical expressions, such as *from my point of view*, from the rest of the sentence. Without the comma, the model selected *of* as the most probable preposition following *because*, instead of *from*. A set of heuristics was used to locate common sites of comma errors and skip these contexts.

There were two other common sources of classification error: antonyms and benefactives. The model very often confused prepositions with opposite meanings (like *with/without* and *from/to*), so when the highest probability preposition was an antonym of the one produced by the writer, we blocked the classifier from marking the usage as an error. Benefactive phrases of the form *for + person/organization* (*for everyone, for my school*) were also difficult for the model to learn, most likely because, as adjuncts, they are free to appear in many different places in a sentence and the preposition is not constrained by its object, resulting in their frequency being divided among many different contexts. When a benefactive appeared in an argument position, the model's most probable preposition was generally not the preposition *for*. In the sentence *They described a part for a kid*, the preposition *of* has a higher probability. The classifier was prevented from marking *for + person/organization* as a usage error in such contexts.

To summarize, the classifier consisted of the ME model plus a program that blocked its application

	Rater 1 vs. Rater 2	Classifier vs. Rater 1	Classifier vs. Rater 2
Agreement	0.926	0.942	0.934
Kappa	0.599	0.365	0.291
Precision	N/A	0.778	0.677
Recall	N/A	0.259	0.205

Table 3: Classifer vs. Rater Statistics

in cases of misspelling, likely punctuation errors, antonymous prepositions, and benefactives. Another difference between the training corpus and the testing corpus was that the latter contained grammatical errors. In the sentence, *This was my first experience about choose friends*, there is a verb error immediately following the preposition. Arguably, the preposition is also wrong since the sequence *about choose* is ill-formed. When the classifier marked the preposition as incorrect in an ungrammatical context, it was credited with correctly detecting a preposition error.

Next, the classifier was tested on the set of 2,000 preposition contexts, with the confidence threshold set at 0.9. Each preposition in these essays was judged for correctness of usage by one or two human raters. The judged rate of occurrence of preposition errors was 0.109 for Rater 1 and 0.098 for Rater 2, i.e., about 1 out of every 10 prepositions was judged to be incorrect. The overall proportion of agreement between Rater1 and Rater 2 was 0.926, and kappa was 0.599.

Table 3 (second column) shows the results for the Classifier vs. Rater 1, using Rater 1 as the gold standard. Note that this is not a blind test of the classifier inasmuch as the classifier's confidence threshold was adjusted to maximize performance on this set. The overall proportion of agreement was 0.942, but kappa was only 0.365 due to the high level of agreement expected by chance, as the Classifier used the response category of "correct" more than 97% of the time. We found similar results when comparing the judgements of the Classifier to Rater 2: agreement was high and kappa was low. In addition, for both raters, precision was much higher than recall. As noted earlier, the table does not include the cases that the classifier skipped due to misspelling, antonymous prepositions, and benefactives.

Both precision and recall are low in these comparisons to the human raters. We are particularly concerned about precision because the feedback that students receive from an automated writing analysis system should, above all, avoid false positives, i.e., marking correct usage as incorrect. We tried to improve precision by adding to the system a naive Bayesian classifier that uses the same features found in Table 1. As expected, its performance is not as good as the ME model (e.g., precision = 0.57 and recall = 0.29 compared to Rater 1 as the gold standard), but when the Bayesian classifier was given a veto over the decision of the ME classifier, overall precision did increase substantially (to 0.88), though with a reduction in recall (to 0.16). To address the problem of low recall, we have targeted another type of ESL preposition error: extraneous prepositions.

7 Prepositions in Prohibited Contexts

Our strategy of training the ME classifier on grammatical, edited text precluded detection of extraneous prepositions as these did not appear in the training corpus. Of the 500-600 errors in the ESL test set, 142 were errors of this type. To identify extraneous preposition errors we devised two rule-based filters which were based on analysis of the development set. Both used POS tags and chunking information.

Plural Quantifier Constructions This filter addresses the second most common extraneous preposition error where the writer has added a preposition in the middle of a plural quantifier construction, for example: *some of people*. This filter works by checking if the target word is preceded by a quantifier (such as "some", "few", or "three"), and if the head noun of the quantifier phrase is plural. Then, if there is no determiner in the phrase, the target word is deemed an extraneous preposition error.

Repeated Prepositions These are cases such as *people can find friends with with the same interests* where a preposition occurs twice in a row. Repeated prepositions were easily screened by checking if the same lexical item and POS tag were used for both words.

These filters address two types of extraneous preposition errors, but there are many other types (for example, subcategorization errors, or errors with prepositions inserted incorrectly in the beginning of a sentence initial phrase). Even though these filters cover just one quarter of the 142 extraneous errors, they did improve precision from 0.778 to 0.796, and recall from 0.259 to 0.304 (comparing to Rater 1).

8 Conclusions and Future Work

We have presented a combined machine learning and rule-based approach that detects preposition errors in ESL essays with precision of 0.80 or higher (0.796 with the ME classifier and Extraneous Preposition filters; and 0.88 with the combined ME and Bayesian classifiers). Our work is novel in that we are the first to report specific performance results for a preposition error detector trained and evaluated on general corpora.

While the training for the ME classifier was done on a separate corpus, and it was this classifier that contributed the most to the high precision, it should be noted that some of the filters were tuned on the evaluation corpus. Currently, we are in the course of annotating additional ESL essays for preposition errors in order to obtain a larger-sized test set.

While most NLP systems are a balancing act between precision and recall, the domain of designing grammatical error detection systems is distinguished in its emphasis on high precision over high recall. Essentially, a false positive, i.e., an instance of an error detection system informing a student that a usage is incorrect when in fact it is indeed correct, must be reduced at the expense of a few genuine errors slipping through the system undetected. Given this, we chose to set the threshold for the system so that it ensures high precision which in turn resulted in a recall figure (0.3) that leaves us much room for improvement. Our plans for future system development include:

1. Using more training data. Even a cursory examination of the training corpus reveals that there are many gaps in the data. Seven million seems like a large number of examples, but the selection of prepositions is highly dependent on the presence of other specific words in the context. Many fairly common combinations of Verb+Preposition+Noun or Noun+Preposition+Noun are simply not attested, even in a sizable corpus. Consistent with this, there is a strong correlation between the relative frequency of a preposition and the classifier's ability to predict its occurrence in edited text. That is, prediction is better for prepositions that have many examples in the training set and worse for those with fewer examples. This suggests the need for much more data.

2. Combining classifiers. Our plan is to use the output of the Bayesian model as an input feature for the ME classifier. We also intend to use other classifiers and let them vote.

3. Using semantic information. The ME model in this study contains no semantic information. One way to extend and improve its coverage might be to include features of verbs and their noun arguments from sources such as FrameNet (http://framenet.icsi.berkeley.edu/), which detail the semantics of the frames in which many English words appear.

References

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

G. Dalgish. 1985. Computer-assisted esl research and courseware development. *Computers and Composition*.

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

National Center for Educational Statistics. 2002. Public school student counts, staff, and graduate counts by state: School year 2000-2001.

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

E. Izumi, 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*.

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

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

M. Murata and H. Ishara. 2004. Three english learner assistance systems using automatic paraphrasing techniques. In *PACLIC 18*.

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