Using an Error-Annotated Learner Corpus to Develop an ESL/EFL Error Correction System

Na-Rae Han, Joel Tetreault, Soo-Hwa Lee, Jin-Young Ha

University of Pittsburgh, Educational Testing Service, Chungdahm Learning, Inc., Kangwon National University
Pittsburgh, PA 15260, USA; Princeton, NJ 08541, USA; Seoul, Korea; Chuncheon, Korea
naraehan@pitt.edu, jtetreault@ets.org, soohlee@chungdahm.com, jyha@kangwon.ac.kr

Abstract

This paper presents research on building a model of grammatical error correction, for preposition errors in particular, in English text produced by language learners. Unlike most previous work which trains a statistical classifier exclusively on well-formed text written by native speakers, we train a classifier on a large-scale, *error-tagged* corpus of English essays written by EFL learners, relying on contextual and grammatical features surrounding preposition usage. First, we show that such a model can achieve high performance values: 93.3% precision and 14.8% recall for error detection and 81.7% precision and 13.2% recall for error detection and correction when tested on preposition replacement errors. Second, we show that this model outperforms models trained on well-edited text produced by native speakers of English. We discuss the implications of our approach in the area of language error modeling and the issues stemming from working with a noisy data set whose error annotations are not exhaustive.

1. Introduction

With the growing adoption of new technologies and computerized applications in language classrooms, applying the latest NLP techniques to the the area of language education is gaining more support. For many ESL (English as a Second Language) and EFL (English as a Foreign Language) students, interacting with computerized applications is an integral part of their learning experience; NLP-based language models can be a valuable tool in assisting teachers and students alike by providing prompt feedback on certain aspects of language, such as mechanical errors, writing quality, and grammatical errors.

Lately there have been efforts aimed at developing grammar correction applications designed specifically with learners of English in mind. A common approach shared by most of the previous work (Izumi et al., 2003; Han et al., 2006; Tetreault and Chodorow, 2008a; Gamon et al., 2008; De Felice and Pulman, 2008) is the reliance on well-formed texts written by native English speakers to train a statistical model. This is mostly due to the fact that to date, constructing a large enough error-annotated corpus to support a statistical approach is time-consuming and labor-intensive. As a result, these approaches train on millions of examples of correct usage and then use a series of thresholds to determine if a writer's usage is correct "enough" given the context. Another issue with this approach is that it fails to model the types of errors and confusions that non-native writers will make.

The main research question we address in this work therefore is: is there an advantage to actually constructing an error-annotated corpus? Specifically, would a model trained on error-annotated data outperform one trained exclusively on well-formed, native text? In this paper, we present a large error-annotated learner corpus, and develop a novel statistical method to ESL/EFL error detection and correction trained exclusively on this corpus. We show that a model trained on examples of correct and incorrect usage, even when the error annotations are not exhaustive, outperforms much larger statistical models trained on native text. In this work, we focus our efforts on preposition error detection and correction since prepositions are among the most difficult for non-native speakers of English to master.

*Please direct all data-related inquiries to Soo-Hwa Lee.

2. Chungdahm English Learner Corpus

We base our model-building experiments on the Chungdahm English Learner Corpus (henceforth Chungdahm Corpus), a collection of English essays written by Korean-speaking students of Chungdahm Institute, a national chain of English language schools run by Chungdahm Learning, Inc., and error-annotated by tutors.[1] The entire data exists in the form of a continuously growing database rather than as a corpus in the strictest sense, but we refer to the portion that we extracted and cleaned up for the purpose of our research as the Chungdahm Corpus. The corpus consists of 131 million words in 861,481 essays for an average essay length of 152 words. The essays were written on 1,545 prompts. There are 6.6 million error annotations made by the tutors on this data set. The specifics of this corpus are shown in Table 1.

There are 45 distinct classes of corrections and feedback, which are categorized under 4 areas: *grammar*, *strategy*, *style* and *substance*. As a part of the grammar-type feedback, a variety of aspects are coached, including spelling, punctuation, verbal forms, subject-verb agreement, noun phrase formulation and prepositions. One notable aspect about the feedback practice is that not all of these diagnostics are applied to every text. The primary role of Chungdahm Institute's essay-writing curriculum as an instructional tool means that only a few selected aspects of English writing are focused on at a particular time and therefore are designated as the target areas for providing feedback. For this reason, it was necessary for the purpose of this study

[1] Students at the lowest 4 proficiency levels are coached by Korean tutors, and the rest at the upper 9 levels are assigned to tutors whose native language is English.

Authors	Students of Chungdahm Institute
Demographic Info	Students of ages 10–16, whose L1 is Korean
Corpus size	130,754,000 words
# of prompts	1,545
Total # of essays	861,481
Avg essay length	152 words
Total # of error annotation	6,605,678

Table 1: Chungdahm Corpus (as of Nov 2008)

to further process the data in order to extract a sub-corpus, which we can be sure were subject to preposition error correction. We discuss this in detail in the next section.

3. The Preposition Data Set

As expected, preposition errors are among the most frequent error types encountered in Chungdahm Corpus: there are 127,345 preposition error annotations, which amount to roughly 2% of all error annotations and other feedback. The prominence of preposition errors in our data is consistent with previous literature on preposition error detection (Chodorow et al., 2007; Tetreault and Chodorow, 2008b). Preposition errors can be categorized into the following three types: *omission*, *commission* (i.e., *extraneous* preposition), and *replacement*, which are illustrated below.

(1) Preposition error types
a. Omission $\langle$NULL, $p\rangle$:
"Yes, I wait $\emptyset$/for you."
b. Commission (extraneous) $\langle p$, NULL$\rangle$:
"So I go to/$\emptyset$ home quickly."
c. Replacement $\langle p1, p2\rangle$:
"Adult give money at/on New Years day."

There are over 50 different prepositions represented in preposition error annotations, either as the original student choice or as the tutor's correction. Among them, 10 prepositions plus the "no preposition" choice were predominant: *about*, *at*, *by*, *for*, *from*, *in*, *of*, *on*, *to*, *with*, and NULL. Together they account for the 99% of student error tokens; the same top 11 types were found to cover over 97% of all tutor correction tokens, with some differences in individual proportions and rankings.

Based on this observation, a decision was made to limit the scope of our preposition error modeling to those 10 prepositions plus "NULL" (for simplicity, this set is henceforth referred to as the "11 prepositions"). It would be ideal, and certainly not impossible, to include all 50+ types of prepositions, but we believe that the practical advantages offered by simplifying the model far outweigh the rather small amount of actual error cases that are discarded. This reduction is consistent with previous work such as (De Felice and Pulman, 2008) and (Gamon et al., 2008) who focused on 9 and 14 prepositions respectively.

Once we filtered out those preposition error annotations involving prepositions other than those 11, the number of error annotations shrunk down to 122,387, or 96.1% of the entire preposition error annotations. We further excluded those cases that are not genuine preposition errors, such as those involving particles ("He carried in/**on**.") and infinitive *to*s mistakenly categorized as prepositions ("He tried NULL/**to** succeed.") with the help of a parser (Klein and Manning, 2003), which eventually left us with 117,665 annotated preposition errors to use in our model building experiments (item c. in Table 4). All 110 possible $\langle student, tutor\rangle$ pairings are represented in the data, whose distribution is shown in Table 2.

	$\langle s, t\rangle$	count	cnt%	cumul%
1	NULL, to	20102	17.08	17.08
2	NULL, in	8885	7.55	24.63
3	in, on	8768	7.45	32.08
4	NULL, at	6075	5.16	37.24
5	NULL, with	4718	4.00	41.25
6	at, in	4648	3.95	45.20
7	to, NULL	4348	3.69	48.90
8	NULL, for	3974	3.37	52.28
9	NULL, of	3956	3.36	55.64
10	in, at	3346	2.84	58.48
11	to, for	2998	2.54	61.03
12	NULL, on	2910	2.47	63.50
13	on, in	2896	2.46	65.97
14	NULL, about	2000	1.69	67.67
15	to, with	1976	1.67	69.34
16	about, with	1896	1.61	70.96
17	for, to	1823	1.54	72.51
18	of, NULL	1755	1.49	74.00
19	of, for	1531	1.30	75.30
20	in, NULL	1466	1.24	76.54
...	...	...	...	...
109	about, by	11	0.0	99.99
110	by, about	4	0.0	100

Table 2: Distribution of $\langle student, tutor\rangle$ preposition correction pairs

The distribution of the 11 prepositions found in either slot of the error annotations is shown in Table 3. One thing immediately noticeable is that the prepositions supplied as corrections by the tutor are more evenly distributed across the 11 categories, whereas the original student prepositions are less so, with a significant portion (46%) concentrated on the NULL choice. As one might expect, pairs involving a NULL student preposition dominate the top ranks, and each pair is represented in a relatively small number.

As explained in the previous section, there are many essays in Chungdahm Corpus that were not coached at all for preposition usage, which we needed to exclude. To achieve this, we compiled a sub-corpus consisting of those essays in which at least one preposition error annotation was found. Admittedly, this simple method has its risks. First, in the process we are sure to lose those essays that were reviewed for preposition errors but were found to be completely error-free, which ideally should to be retained. Secondly, what we do here amounts to manipulating the volume and kinds of well-formed preposition usages to be

prep	student		tutor	
	count	count%	count	cnt%
NULL	54259	46.11	9791	8.32
about	3265	2.77	2883	2.45
at	8556	7.27	11408	9.69
by	1194	1.01	1498	1.27
for	4852	4.12	11110	9.44
from	1085	0.92	4099	3.48
in	16705	14.19	19846	16.86
of	6328	5.37	5919	5.03
on	4577	3.88	15062	12.80
to	14497	12.32	25047	21.28
with	2347	1.99	11002	9.35
	117665	100%	117665	100%

Table 3: Distribution of 11 prepositions in $\langle student, tutor \rangle$ correction pairs

a.	text size (words)	20,472,948
b.	# of essays	111,060
c.	# of all preposition error tokens (c_1+c_2+c_3)	117,665
c_1.	# of omission $\langle$NULL, $p\rangle$ error tokens	54,259
c_2.	# of extraneous $\langle p$, NULL$\rangle$ error tokens	9,791
c_3.	# of replacement $\langle p1, p2\rangle$ error tokens	53,615
d.	# of preposition tokens with no error annotation	1,104,752
e.	# of all preposition tokens (c+d)	1,222,417
f.	preposition error rate in data (c/(c+d))	9.6%

Table 4: Data set used for preposition modeling

introduced into our experiments, which has direct ramifications in the resultant models.

Even with this filtering, we later learned that many overlooked and unannotated errors still exist in this data set, which indicates error annotation is not applied exhaustively within a text. We will discuss later in Sections 5. and 7.2. how our approaches have direct ramifications in the resultant models.

In the end, the selected subset of the Chungdahm Corpus to be used in our experiments consisted of 111,060 essays and 20,472,948 words which yielded 1.2M preposition tokens. Specifics of this sub-corpus are shown in Table 4.

4. A Maximum-Entropy-Based Model for Preposition Prediction

In our language model, a preposition use is represented as an ordered pair $\langle s, c\rangle$ where s indicates the original, potentially incorrect, preposition choice made by the student, and c the correct preposition. s and c range over the set of 11 preposition types {NULL, *about*, *at*, *by*, *for*, *from*, *in*, *of*, *on*, *to*, *with*}. In preposition cases where there is no error annotation, the original student choice is presumed correct and c defaults to s. Our goal is to build a classifier whose choice of outcome m matches c.

Text & Annotation:

snow	is	falling	there	*at*	the	winter	.
	-3	-2	-1	*s*	+1	+2	+3
		MOD				ARG	
$\langle s, c\rangle$:	$\langle at, in\rangle$						

Event:

outcome: *in*	
features:	
name	value
s	at
wd_{-1}	there
wd_{+1}	the
MOD	falling
ARG	winter
MOD_ARG	falling_winter
MOD_s	falling_at
s_ARG	at_winter
MOD_s_ARG	falling_at_winter
$wd_{-1,2}$_s	falling_there_at
s_$wd_{+1,2}$	at_the_winter
3GRAM	there_at_the
5GRAM	falling_there_at_the _winter
wd_L	snow, is, falling
wd_R	the, winter, .
MODt_ARGt	VBG_NN
MODt_s	VBG_at
s_ARGt	at_NN
MODt_s_ARGt	VBG_at_NN
TRIGRAMt	NN_at_DT

Table 5: Event representation of a preposition token

To this end, we employed a maximum entropy (ME) model (Ratnaparkhi, 1998) as our machine learning method of choice. Maximum entropy has been shown to perform well in combining heterogeneous forms of linguistic evidence. Out of the entire set of 1.2M preposition tokens, approximately 200K were set aside for development purposes, and the remaining set of about 978,000 preposition tokens was used in training a ME classifier. This is a relatively small training data size; previous work in preposition error detection has typically used training models built from millions of preposition events.[2]

In the model, each proposition token along with its context is represented as a preposition *event* (Table 5): a preposition event consists of an *outcome*, the correct preposition choice, and a set of contextual *features*, each of which encoding a particular aspect of the linguistic context surrounding the preposition instance, including the original

[2]For example, Tetreault and Chodorow (2008a) used a training set composed of 7 million events.

Code	*Description*	*Example*
OK	Writer's preposition is acceptable	"So she has a lot *of* money."
Wrong Choice	Writer used the wrong preposition	"So she got married *with* him."
Extraneous Use	Writer used a preposition in a context that does not license one	"Next *to* year, Jennifer and Mike are married."
Indecipherable	Context is too messy or confusing to make a judgment on preposition usage	"Thry thought museumis fit *to* knowreal world."

Table 6: Evaluation corpus preposition annotation scheme

preposition choice made by the student writer.

Designing a set of contextual features that are strong predictors of a particular preposition outcome is paramount in achieving good system performance. Some of these features are based on surface phenomena such as nearby word forms; others require more sophisticated linguistic knowledge such as the part-of-speech of a word or the two elements that perhaps play the most definitive role in preposition choice: the lexical head of the phrase which the preposition modifies (MOD) and the lexical head of the preposition argument noun phrase (ARG). We used the Stanford Parser (Klein and Manning, 2003) to identify the MOD and ARG. The following configurations are targeted in the local context of a preposition: (a) s the preposition token chosen by the student, (b) wd_{-1} the word immediately preceding s, (c) wd_{+1} the word immediately following s, (d) MOD, (e) ARG, (f) 3 words preceding s, (g) 3 words following s. Actual features are built from these bits of information by combining related ones (e.g., "trigrams" concatenating (a), (b), (c)) and/or substituting part-of-speech tags for actual lexical entries (e.g., POS trigrams). Table 5 illustrates the entire feature set used in our experiments for the sample sentence "Snow is falling there *at* the winter." which has the correction of *in*.

In extracting features, we applied limited semantic processing in order to generalize on items of open lexical categories, such as numbers and person names. Specifically, digits and numbers were collapsed into representative forms (e.g., *1987* → *1111*, *thirty-five* → *eleven*), and so were hyphenated Korean names (e.g., *Min-kyoung*, *Su-Hee* → HYPHEN-NAME).

5. Evaluation

5.1. Evaluation Corpus

Because the error flagging in the Chungdahm Corpus is far from exhaustive, automated methods of evaluation could not be relied upon beyond its utility in the feature selection process[3]. For a true measure of system performance, therefore, we conducted a round of manual annotation to create a fully error-annotated evaluation set from a subset of the corpus.

To create this set, three trained raters annotated 1,000 preposition contexts randomly selected from the set-aside portion of our data set. The original annotations were not presented to them. The raters followed the annotation scheme presented in Tetreault and Chodorow (2008b) in which the writer's preposition is rated on a 4-point scale: (1) OK: writer's preposition is acceptable, (2) Wrong Choice: writer used the wrong preposition, (3) Extraneous Use: writer used a preposition in a context that does not license one, and (4) Indecipherable: context is too messy or confusing to make a judgment on preposition usage. See Table 6 for examples of each of the four categories. All three raters also judged an overlap set of 100 preposition contexts to compute kappa. Agreement ranged from 0.860 to 0.910 and kappa from 0.662 to 0.804, which are on par with those reported in Tetreault and Chodorow (2008b).

Next, we compared the new annotations in our evaluation corpus with the original ones supplied by Chungdahm tutors. Agreement and kappa between the two annotations were 0.827 and 0.426 respectively. More importantly, the comparison revealed that indeed many genuine preposition errors are left unflagged in Chungdahm Corpus: 57.4% of all replacement-type errors and 85% of extraneous preposition errors were found to be unmarked in the original annotation supplied by Chungdahm tutors. This problem of incomplete error annotation has direct consequences for the resulting models, which is discussed in detail in Section 7.2.

5.2. Evaluation of Learner Model

Our system works primarily as a multi-outcome prediction model (an 11-way classifier), which produces a prediction on the *correct* preposition choice given a context: for a preposition use represented as $\langle s, c\rangle$ (where s indicates the original, potentially incorrect, preposition choice made by the student, and c the correct preposition), the machine's choice of outcome m purports to match c. When used as an error diagnostic tool, it not only detects the existence of a preposition error when the model prediction differs from the student's original choice (i.e., when $m \neq s$) but also supplies correction in the form of the model's preposition choice m. In other words, the model produces a multi-outcome decision (error detection *and* correction: "The model suggests preposition m as the correct alternative to the student choice s") which can be backed off to a binary decision (error detection: "The model's preposition prediction m differs from the student choice s")[4]. The former is successful iff $s \neq c$ and $m = c$; the latter is successful iff

[3] Automated evaluation, relying on the development portion of Chungdahm Corpus, was used as a basis for verifying that addition and/or removal of certain features lead to a performance gain or loss.

[4] An exception to this is the omission error type $\langle \text{NULL}, p\rangle$, which assumes an input that has already been resolved for error detection, as we shall see shortly.

$s \neq c$.

Since the model's prediction is ultimately used as a diagnostic tool, its performance is measured in two key figures: *precision* ("Of all error flags that the system raises, how many of them are correct?") and *recall* ("Of all existing errors, how many does the system successfully diagnose?). The performance of our model is reported separately for the three types of preposition errors in Table 7.

$\langle$NULL, $p\rangle$	accuracy
error correction	0.833

$\langle p$, NULL$\rangle$	precision	recall
error detection	1	0.049
+ correction	0.87	0.043

$\langle p1, p2\rangle$	precision	recall
error detection	0.933	0.148
+ correction	0.817	0.132

Table 7: Performance of omission $\langle$NULL, $p\rangle$, extraneous $\langle p$, NULL$\rangle$, and replacement $\langle p1, p2\rangle$ type errors

For the replacement and extraneous error types (e.g., $\langle in, on\rangle$ and $\langle to$, NULL$\rangle$), both precision and recall figures are presented for the two types of decisions. For omission type errors (NULL as the student preposition, $\langle$NULL, $p\rangle$), the "error vs. non-error" binary decision has effectively been made: the classifier learns to predict that given NULL as the original student choice the correct preposition has to be an overt one ($x \neq$ NULL for $\langle$NULL, $x\rangle$). The accuracy of the model's alternative preposition suggestion is, therefore, the only relevant performance measure for this type. This means that as far as omission errors are concerned, our model does not actually identify them in a novel text; it must depend on the output of another model whose specific task is identifying missing prepositions, as is done in Gamon et al. (2008), for which it is then able to recommend the correct proposition choice. We have plans to implement such a model in the future.

Overall, the results indicate good levels of precision though low recall. The model's intended use as an instructional tool, however, means more emphasis is placed on precision than recall; the goal of reducing "false positives", that is, a system diagnosing an error where the student choice is in fact correct, is paramount. Furthermore, upon close inspection it was found that none of the false positive cases of error detection actually involved those preposition usages ruled "OK" by annotators; they were all of the "Indecipherable" type. Therefore, the system's "false positives" are not genuine cases of the system erroneously flagging grammatical preposition usages as ungrammatical. In this regard, we believe our system achieves a performance level that is suitable for operational use, albeit with ample room for improvement in the recall rates. We believe that our model's low recall rates are directly related to the non-comprehensive error annotation in the training data; again see Section 7.2. for further discussion.

5.3. Comparison with Models Trained on Native Text

One of the goals of our study is to provide a comparison between a learner-corpus-based model and a model trained on well-formed, native text (i.e., text produced by native speakers of English). To accomplish this, we trained a statistical model on texts from the Lexile Corpus, a collection of K-12 reading materials. As one might immediately notice, this comparison is inherently biased, as we are weighing between a model trained and tested on the same material (the learner model) and one trained on one set of data and then tested on another (the native model). It should be noted, however, that the point of this exercise is not about obtaining a completely balanced comparison between the two models; it is meant to provide a comparison point between the two general *approaches* to L2 learner error modeling, namely, our current approach of relying on error-corrected learner language and the other, more prevalent, one that relies entirely on native-speaker corpora.

In addition, we took further steps to minimize the disparity between the two corpora; a decision was made to limit our data set to the texts from the 7th and 8th grade reading levels of the Lexile Corpus, which were deemed to have the closest writing style and content as those of the Chungdahm Corpus, with the exclusion of all other, lower and higher, grade reading levels. While some might argue that there is still too big a disparity between the Lexile data and the Chungdahm corpus, we would like to point out that the training corpora employed by previous studies are even more dissimilar to the typical texts produced by L2 learners (San Jose Mercury News data were used in Chodorow et al. (2007), Tetreault and Chodorow (2008a), Tetreault and Chodorow (2008b) along with the Lexile Corpus; Microsoft Encarta Encyclopedia and Reuters News were used in Gamon et al. (2008)).

One advantage native texts have over error-annotated learner corpora is the fact that they are available in a much larger size. In order to take advantage of this, we trained five differently sized models ranging from 1 million, roughly the size of our learner model, up to 5 million in training event size[5]. When tested on a held-out portion of the same data set (thus training and testing on well-formed native text), these models show performance levels comparable to what is reported in previous research for the same task (Tetreault and Chodorow, 2008a; Gamon et al., 2008; De Felice and Pulman, 2008): the accuracies of multi-outcome preposition classification task for the four models ranged from 0.694 (the 1 million model) to 0.740 (5 million).

An identical set of features are used for these models except for those that include references to the original preposition choice, which were either dropped altogether or altered to remove such references. Note that the students' preposition

[5]It is certainly possible to construct a native-speaker-produced corpus which is at least one or two order of magnitudes larger. It would, however, necessitate inclusion of a large amount of text whose writing style and substance differ vastly from those represented in Chungdahm Corpus. To ensure the genre compatibility between the two corpora to the best of our ability, we limited our comparison corpus to the sub-section of the Lexile Corpus, as noted above.

choice does play a role in the application of the native-text-trained models as well, but in a different stage. For our learner-data trained model, the student preposition choice is built into the training data and therefore informs the model itself; for models trained on native text, the crucial piece of information is consulted in the diagnosis-producing phase, when *thresholds* are applied.

For models trained on native text, it is essential that thresholds are set so that the models are allowed skip error diagnosis on those cases with lower confidence. With no such mechanism in place, these models typically over-diagnose errors, resulting in high recall but low precision. In an approach consistent with those used in similar systems (Tetreault and Chodorow, 2008a; Gamon et al., 2008; De Felice and Pulman, 2008), we let the models skip diagnosis on those cases where: (1) the difference between the probabilities of the two top preposition choices is less than 0.8 or (2) the probability of the student preposition choice is greater than 0.1. Also consistent with previous work, we favored precision over recall. This tactic was taken because the goal is to reduce the number of false positives, that is, reducing the cases where the system diagnoses an error where the student choice is in fact correct.

$\langle p1, p2 \rangle$	error detection		+ correction	
	precis.	recall	precis.	recall
Learner	**0.933**	**0.148**	**0.817**	**0.132**
N-1mil	0.536	0.132	0.416	0.106
N-2mil	0.585	0.142	0.463	0.116
N-3mil	0.594	0.126	0.453	0.099
N-4mil	0.583	0.153	0.462	0.125
N-5mil	0.605	0.147	0.484	0.121

Table 8: Performance comparison on replacement $\langle p1, p2 \rangle$ preposition errors

Table 8 shows comparison results for replacement type errors $\langle p1, p2 \rangle$. The native-trained models, by design, are unable to handle the other two types, which involve NULL preposition choices. The main finding is that the Learner model outperforms all of the native-trained models, even the N-5mil model which was trained on a data set that is five times as large. Overall, the recall for all models is low, but the precision of the Learner model far exceeds that of the native-trained models.

From this, we conclude that models trained on error-annotated learner corpora, albeit with inexhaustive error annotation, not only have a competitive advantage over similarly sized native corpora but also much larger sized ones as well.

6. Related Work

There have been a few previous efforts aimed at building an application for diagnosing preposition errors by English learners. Not all of them, however, include evaluation on genuine, learner-produced text, opting instead for testing on native texts only (Lee and Seneff, 2006; De Felice and Pulman, 2007). We review here those ones that do.

Chodorow et al. (2007) designed a preposition error detection model targeting as many as 34 prepositions. They trained a ME-based binary classifier on 7 million preposition events extracted from a large corpus of native texts and applied it to 2,000 preposition cases from non-native texts. They report 0.88 overall precision and 0.16 recall on detecting replacement type preposition errors, and 0.796 precision and 0.304 recall for replacement and extraneous type errors. In a follow-up work (Tetreault and Chodorow, 2008a), they experimented with combination features and incorporating data from additional corpora, and report precision and recall figures of 0.84 and 0.19.

Gamon et al. (2008), on the other hand, targeted all three types of preposition errors involving 14 prepositions (not including NULL), from identification to correction. They employed two classifiers, one to determine whether a preposition/article should be present and one for the correct choice, and an additional model as a filter. Their system was trained on large sets of native text and tested on 8,000 English sentences by Chinese learners of English. Its 362 system error diagnoses on preposition use were then judged by a human reviewer: precision was about 80% and recall was not reported.

Izumi et al. (2003) and Izumi et al. (2004) are the only previous work that we are aware of that employ an approach similar to our own, namely, training and testing solely on learner data. Their data size, however, is rather too small: using the Standard Speaking Test Corpus as their source, the core of their data set consists of English interview transcripts of 56 Japanese speakers totaling 6,216 sentences. They do not present performance for prepositions specifically, but overall performance for all of the 13 grammatical error types they targeted was at 25% precision and 7% recall.

More recently, Tetreault and Chodorow (2009) used region web counts to discover typical preposition replacement errors made by different language groups. They showed that a statistical classifier trained on well-formed text could be improved by augmenting it with information about preposition constructions that are problematic for non-native English speakers. Their work is similar in spirit to ours in that they are leveraging data about errors that non-native speakers typically make.

7. Discussion

7.1. Learner Language vs. Native Corpora

As mentioned earlier, the central methodology that is common to much previous research is use of native-produced texts as the basis of statistical modeling. Essentially, this method creates a model of well-formed, native speakers' English, and uses it to make a prediction on learner language, which, when found to outweigh the learner-produced choice, is turned into an error diagnosis.

The motivation behind this method is not so much one from theoretic considerations on its robustness as a practical one: the simple fact that well-edited, large-scale English corpora are readily available resources. The main approach in our work is distinct in this regard: we train a model of L2 (second language) English correction entirely on L2 output and its error annotation. In a way, it constitutes building a direct model of the L2.

In particular, the language error model presented in this study is that of the students of Chungdahm Institute, that is, 10-16 year old English learners whose L1 (first language) is Korean. This, then, ultimately raises the question of the extensibility of our model to other varieties of learner language, for instance, the English of Chinese college students, the English of French-born residents in the US, or even the English of Korean adults. Arguably, there is a certain universal quality to the native-corpora-based approach, in which a single model of idealized English is created, which just might prove it better-suited for wider application. This prognosis, of course, is something that needs empirical verification. Furthermore, it would also be interesting to investigate possible ways to combine the two approaches in an attempt to reap the benefits that each of them has to offer. One potential method is to have a system with specific models for different L1s and then a generic model to which it can back-off when it does not have a model or the L1 model has a low confidence in its decision.

7.2. The Problem of Partial Error Annotation

Though providing a unique and invaluable resource for second language error modeling, the Chungdahm English Learner Corpus has its weaknesses, of which the problem of partial error annotation challenge. As noted earlier in Section 5.1., we estimate that only about 43% of the replacement errors and 15% of the extraneous preposition errors are corrected by tutors.

On a practical level, it renders cumbersome the evaluation process of a system developed on it, as we saw in Section 5.. Reliance on human evaluation means quick modification and testing of the system for either performance gain or other testing purposes is no longer feasible. One can either extrapolate from automated evaluation results to roughly gauge the system's true performance, or alternatively create a set of testing data with accurate and thorough error annotations applied, as we did here, which however will have to be fairly large in order for the diverse patterns and contexts of preposition misusages to be represented in it.

Another, more fundamental, kind of question is the problem of training a model on a set of data that includes conflicting evidence. In our data set, evidence of an error is valid, assuming, of course, a perfect accuracy of those error corrections that are present[6], while evidence of a non-error may or may not be. The low recall rates of our trained model are the direct consequences of this: the system not only assumes a lower-than-true error rate, which overwhelmingly favors the original preposition choice by the student, but it also has to work with the disadvantage of having to reconcile between conflicting sets of evidence.

As much as we would like our data set to be perfect for our projects, the issue of partial error annotation is inherent in the provenance of derived data such as ours: they are a byproduct of an instructional courseware, and providing full and thorough error correction for student errors is not in the interest of the educational institutions who initiate it. Meanwhile, efforts to create error-annotated learner corpora for purely academic and research purposes have so far engendered smaller corpora, much too small for machine learning methods, although presumably with more exhaustive and principled sets of annotations (Chinese Learners' English Corpus, 1 million words; the Standard Speaking Test Corpus, 1M). Even with the few large-scale corpora available, error tagging is done only on a subsection of the data (Longman Learners' Corpus, 10M; HKUST Corpus, 25M), or the corpus is not publicly available (The Cambridge Learner Corpus, 20M) (Xiao, 2008).

Given the alternatives, and also the promising early results presented in this study, partially error-annotated large-scale data sets are attractive resources to be exploited for statistical modeling of learner language. It is, then, critical for the researchers to engineer solutions that address the challenges posed by incomplete error annotation. One potential approach that might prove effective for our data would be to apply bootstrapping methods; through successive training sessions, those non-error cases whose validity is deemed highly suspect can be re-labeled for next runs. We plan on pursuing this avenue in future work.

8. Conclusions

In this paper we investigated the impact of using a vast error-annotated data set for the tasks of ESL/EFL error detection and correction. Our results showed that even with a partially error-annotated set, a model that leveraged the corrections drastically outperformed models of the same size trained on well-formed native text, as well as models five times the size. We believe this shows how much one can expect to increase performance in a statistical system by leveraging such large error-annotated corpora, though we do acknowledge the non-trivial time and cost expenses. Also of note is that this work shows that an exhaustive annotation is not necessary to outperform a standard native-trained model. This note is of significance because it has implications for expediting annotation procedures by utilizing existing resources with noisy annotation.

In sum, we have presented a method of building an error identification and correction model of preposition usage based on English texts that are produced by L2 learners and partially annotated for errors. A first attempt at training such a system exclusively on a large set of learner texts, our approach shows that such a method is not only viable but also leads to good system performance. For future work, we are planning to experiment with larger native-trained models to investigate how much is required for a native-trained model to approach the performance of a learner-trained model. In addition, we will investigate if a model trained on Korean learner data can be effective in detecting errors by writers of other L1s.

9. Acknowledgments

We are grateful to Chungdahm Learning, Inc. for making their valuable learner essay data available for this research. We would also like to thank the members of the Strategic R&D Center for their support in this project, anonymous reviewers for their valuable feedback, and our annotators who helped with evaluation.

[6]Consistency and accuracy of error annotation have been shown to be difficult to achieve; detailed discussions can be found in Tetreault and Chodorow (2008b).

10. References

Martin Chodorow, Joel R. Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30.

Rachele De Felice and Stephen G Pulman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the ACL-07 Workshop on Prepositions*.

Rachele De Felice and Stephen G. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 english. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexander Klementiev, William Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2004. Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129.

George Heidorn. 2000. Intelligent writing assistance. In Robert Dale, Herman Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 181–207. Marcel Dekker, Inc.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 145–148, Morristown, NJ, USA. Association for Computational Linguistics.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. SST speech corpus of Japanese learners' English and automatic detection of learners' errors. *ICAME Journal*, 28:31–48. http://icame.uib.no/ij28/.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–230.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *AAAI '94: Proceedings of the Twelfth National Conference on Artificial Intelligence (vol. 1)*, pages 779–784.

John Lee and Stephanie Seneff. 2006. Automatic grammar correction for second-language learners. In *Proceedings of Interspeech 2006*, Pittsburgh, PA.

John Lee and Stephanie Seneff. 2008. Correcting misuse of verb forms. In *Proceedings of ACL 2008*, Columbus, Ohio.

Ting Liu, Mingh Zhou, Jianfeng Gao, Endong Xun, and Changning Huan. 2000. PENS: A machine-aided English writing system for Chinese users. In *Proceedings of ACL 2000*, pages 529–536.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

Hisami Suzuki and Kristina Toutanova. 2006. Learning to predict case markers in Japanese. In *Proceedings of COLING-ACL*, pages 1049–1056.

Joel Tetreault and Martin Chodorow. 2008a. The ups and downs of preposition error detection. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK.

Joel R. Tetreault and Martin Chodorow. 2008b. Native judgments of non-native usage: Experiments in preposition error detection. In *COLING Workshop on Human Judgments in Computational Linguistics*.

Joel Tetreault and Martin Chodorow. 2009. Examining the use of region web counts for ESL error detection. In *Proceedings of Web as Corpus Workshop (WAC-5)*, San Sebastian, Spain.

Jenine Turner and Eugene Charniak. 2007. Language modeling for determiner selection. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational LinguisticsLinguistics; Companion Volume, Short Papers*, pages 177–180.

Richard Xiao. 2008. Well-known and influential corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*, Handbooks of Linguistics and Communication Science. Mouton de Gruyter, Berlin.

Xing Yi, Jianfeng Gao, and William B. Dolan. 2008. Web-based english proofing system for English as a second language users. In *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*.