Using Statistical Techniques and Web Search to Correct ESL Errors

Michael Gamon
Microsoft Research

Claudia Leacock
Butler Hill Group

Chris Brockett
William B. Dolan
Jianfeng Gao
Dmitriy Belenko
Microsoft Research

Alexandre Klementiev
University of Illinois at Urbana Champaign

ABSTRACT

In this paper we present a system for automatic correction of errors made by learners of English. The system has two novel aspects. First, machine-learned classifiers trained on large amounts of native data and a very large language model are combined to optimize the precision of suggested corrections. Second, the user can access real-life web examples of both their original formulation and the suggested correction. We discuss technical details of the system, including the choice of classifier, feature sets, and language model. We also present results from an evaluation of the system on a set of corpora. We perform an automatic evaluation on native English data and a detailed manual analysis of performance on three corpora of nonnative writing: the Chinese Learners' of English Corpus (CLEC) and two corpora of web and email writing.

KEYWORDS

Computational Linguistics, Automatic Error Detection, Data-Driven Error Detection

INTRODUCTION

Automated identification and correction of errors made by learners of English as a second or foreign language present special challenges to computational linguists. Conventional grammar-proofing tools typically target errors made by native speakers and leave common classes of learner errors—errors that native writers rarely make—unchecked. In addition, the more error-prone input of nonnative speakers often prevents parser-based analyzers from successfully identifying potential problems in nonnative writing.

Using a combination of statistical and rule-based techniques, we have developed the Microsoft Research (MSR) ESL Assistant, a prototype that targets common classes of errors made by native speakers of East Asian languages (Mandarin Chinese and other Chinese "dialects," Japanese, and Korean). A unique component of the tool enables the language learner to access examples on the web to determine whether the suggestion is appropriate to their intent. In this paper we present an outline of the system and an evaluation of its performance on four data sets.

TYPICAL ERRORS MADE BY CHINESE AND JAPANESE LEARNERS

Although existing rule-based proofing tools, such as the grammar checker incorporated into Microsoft Word™ and other Microsoft Office™ products (Heidorn, 2000), do not target nonnative speakers, they do in practice quite robustly address certain writing errors (e.g., subject-verb disagreement) that are widely observed among English learners irrespective of their L1. In the interests of not reinventing the wheel, we sought to identify classes of errors that are observable with frequency in English writing by nonnative writers but which are not specifically addressed by such tools. We therefore informed our priorities by drawing on two significant tagged corpora of learner errors that have recently become available in China and Japan, two countries that are of natural commercial interest, and which are believed to share "sprachbund"-like characteristics with respect to learner error patterns: The Chinese Learner English Corpus (Gui & Yang, 2003; see Gui & Yang, 2001 for an English language description) and the National Institute of Information and Communications Technology (NICT) Japanese Learner English (JLE) corpus (Izumi, Uchimoto, & Isahara, 2004, 2005).

The Chinese Learner English Corpus (CLEC) is a 1-million-word corpus of short essays written by Chinese high school and university students at different skill levels. Gui and Yang (2001) report that error locations and error types in this corpus were tagged using machine-aided manual tagging. This is a characteristic that militates against its direct application as a training/testing corpus for error correction involving machine-learning solutions since casual investigation reveals that large numbers of errors remain uncaptured by their automated error detection techniques and follow-up human validation has not been altogether reliable. A huge portion of the tagged errors in this corpus involves lexical errors of a sporadic and often highly idiosyncratic nature that do not appear readily amenable to solutions. Moreover, many of the error categories are obscure; for example, a high proportion of collocation and verb subcategorization errors turn out to fall within the broad class of preposition errors. Despite these difficulties, the writing samples themselves present a valuable source of error data for manual evaluation. In addition, the detailed breakdown of the errors provided by Gui and Yang afford useful insights into the distribution of error types with potential to help shape both pedagogical strategies and learning and writing aids for Chinese learners of English.

The NICT JLE corpus comprises transcriptions of 300 hours of oral interviews conducted with Japanese students. A subset of this 1.2 million-word corpus was manually tagged for 47 error types by a native speaker of English. Since the corpus is based on spoken interactions that were elicited in a potentially stressful interview situation, the errors do not necessarily represent those that might occur under the more deliberative conditions that obtain in writing where self-correction is possible.

We found that the error tags in both these corpora require interpretation, owing to the overall sparseness of, and ambiguity among, the error classes that they propose. For our purposes, we discounted nonsystematic lexical errors in the CLEC corpus, such as lexical confusions or simply incomprehensible word choice. An example of such a nonsystematic error is the sentence *Our class is an emphatic class* where *emphatic* may have been confused with *empathetic*. Similarly, we discounted speech-related errors (e.g., self-corrections and hesitations) in the JLE corpus. Our goal was to identify broad error types at a level of frequency beyond the anecdotal that might warrant initial attention. Of these, two high-incidence classes—articles and prepositions—stood out as warranting particular focus: successfully accounting for just these two classes (assuming perfect precision and recall) would permit coverage of approximately one third of all nonspeech-related errors tagged in the JLE corpus.

Our analysis of the JLE corpus indicates that of the 9,173 tagged errors that were not speech related, 26.60% involved article errors. Failure to insert a required article accounted for 70.37% of article errors, while superfluous insertion accounted for a further 14.34%. In other words, approximately 85% of the article errors relate primarily to whether or not to insert an article, the remainder being associated with choice of article, suggesting that the system might benefit more by focusing on the former rather than the latter. Approximately 10% of the CLEC nonlexical errors appear to be article or noun-number related, although differences in the corpus type (speech vs. written) and other factors make further comparison difficult.

In the same vein, approximately 10% of JLE nonspeech errors involve prepositions. The JLE data set distinguishes three classes of preposition error: "ordinary" prepositions not subcategorized for by a lexical item, prepositions in complements of verbs or adjectives, and errors relating to the complements of prepositions. Ordinary preposition errors are the most frequent error type (a total of 553 error tokens), and among these, 55.34% involve missing prepositions, while 13.02% involve insertion of prepositions in inappropriate locations. In the case of errors involving a preposition in the complement of a verb or adjective (312 instances), omissions account for 68.91%, while only 7.37% of errors involve insertion of a preposition where none is needed. Errors involving the complements of prepositions are few (only 33 cases), predominantly involve use of gerunds, and probably constitute a separate class. In the CLEC data, only 2% of nonlexical errors are identified as preposition related, but the actual percentage is probably substantially higher since many preposition errors in this corpus appear to be subsumed under collocation and verb complement selection error tags.

Given these error distributions, the primary targets of the ESL Assistant system are as follows, together with their approximate adjusted frequencies in the JLE corpus:

1. definite and indefinite article presence and choice (JLE: 26.60%)
   We should think whether we **have ability** to do it well.

2. preposition presence and choice (JLE: 10.40%)
   Finally, the pollution **on** the world is serious.

3. noun number (JLE: 8.26%)
   So other **works** couldn't be done in adequate **times**.

4. gerund/infinitive confusion (JLE: 3.02% of errors involve verb or adjective complement structure)
   So, money is also important in **improve** people's spirit.

5. auxiliary verb presence and choice (JLE: 6.74% are verb choice errors, the overwhelming majority of instances being auxiliary related)
   The fire will break out, it can **do harmful** to people.

6. adjective/noun confusion (JLE: 3.19%)
   There was a wonderful women volleyball match between Chinese team and **Cuba** team.

7. local word order, for example, adjective sequences and nominal compounds (JLE: 3.19%)
   A **pop British band** called "Spice Girl" has sung a song.

8. overregularized verb inflection (These account for only 0.13% of JLE errors but are usually not correctly covered in the Word spell checker)
   It was **builded** in 1995.

DIFFERENT SOLUTIONS FOR DIFFERENT TYPES OF ERRORS

Previous approaches to language learner error correction fall into two basic categories: rule-based approaches and data-driven approaches. Eeg-Olofsson and Knutsson (2003) report on a rule-based system that detects and corrects preposition errors in Swedish texts produced by nonnative writers. Rule-based approaches have also been used to predict definiteness and indefiniteness of Japanese noun phrases as a preprocessing step for machine translation into English (Murata & Nagao, 1993; Bond, Ogura, & Ikehara, 1994; Heine, 1998), a task that is similar to the prediction of English articles. Data-driven approaches have gained popularity throughout the past decade and have been applied to article prediction in English (Knight & Chander, 1994; Minnen, Bond, & Copestake, 2000; Turner & Charniak, 2007), an array of Japanese learners' errors in English (Izumi, Uchimoto, Saiga, Supnithi, & Isahara, 2003), and article and preposition correction in ESL text (Han, Chodorow, & Leacock, 2004; Nagata, Wakana, Masui, Kawai, & Isu, 2005; Nagata, Kawai, Morihiro, & Isu, 2006; De Felice & Pulman, 2007; Chodorow, Tetreault, & Han, 2007; Tetreault & Chodorow, 2008a; Gamon et al., 2008).

The errors we described above exhibit very heterogeneous properties in terms of the complexity of their solutions. Preposition and article errors are at one extreme of the spectrum: large amounts of contextual information are necessary in order to arrive at a correction. On the other extreme, overregularized verb inflection is detectable without any contextual information: the form *writed* is simply one of a very limited set of forms that result in overregularization of one of the 100+ irregular verbs of English.

In our system we decided to take these very different error properties into account by choosing different techniques for different error types. The contextual information that is needed for preposition and article correction, for example, lends itself to a machine-learning approach that derives generalizations about preposition and article use from large amounts of training data. Overregularized inflection, on the other hand, can be targeted by a lookup rule and a list of overregularized forms of irregular verbs.

Given that different techniques are used for different error types, we opted for a modular design in which each error type is targeted by an autonomous module. In the final design of our system, four modules are machine learned: preposition presence/choice, article presence/choice, gerund/infinitive confusion, and auxiliary verb presence/choice. Nineteen other modules are heuristic in nature, requiring regular expressions to match textual patterns and small sets of local features. The modular design has the added advantage of allowing easy extension of the targeted error types by simply adding new modules.

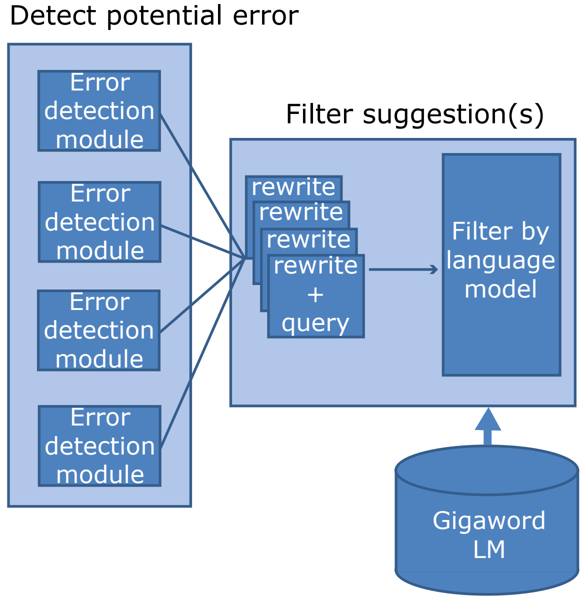
The remainder of this article focuses primarily on the machine-learned modules for article and preposition errors because they are the most frequently occurring error types for nonnative English language learners of East Asian background as well as being difficult to resolve automatically.

ESL ASSISTANT SYSTEM OVERVIEW

Each of the error-specific modules provide the initial error detection and correction suggestions. They all have access to two sources of information: the original user input and part-of-speech tags assigned to each token in the original input by a part-of-speech tagger (Toutanova, Klein, Manning, & Singer, 2003). In addition, individual modules have access to feature sets that are relevant to that error type. However, using this set of modules alone overgenerates corrections, a situation that is already difficult and potentially annoying enough for a native speaker to deal with, let alone a nonnative speaker. In order to further restrict and improve the final set of suggested corrections, we added a filter that is based on a very large language model. This language model is trained on the Gigaword corpus (Linguistic Data Consortium, 2003) and utilizes 7-grams with absolute discount smoothing (Gao, Goodman, & Miao, 2001; Nguyen, Gao, & Mahajan, 2007). For each detected error and suggested correction that is produced by the error modules, the language model provides a score for both the original (uncorrected) user input and the corrected version. Only when the corrected version achieves a substantially higher language model score than the original user input do we surface the suggested correction to the user.

Previous experiments (Gamon et al., 2008) indicate that the reduction of suggested corrections on native text achieved by the language model filtering stage amounts to 67% on the preposition suggestions and 51% on the article suggestions, accompanied by a steep increase in precision, albeit at the cost of recall. At the same time, the language model itself—without any error-specific classification stage—would not be able to predict errors with sufficient accuracy, achieving only 58.36% accuracy in the preposition task on native data compared to 77.55% achieved by the preposition classifier. On nonnative data, where the goal is to maximize precision, the language model allows us to carefully threshold the suggested corrections in order to minimize the number of false suggestions. The overall design of ESL Assistant is illustrated in Figure 1.

Figure 1
Error-Specific Modules and Filtering by Language Model Score



The four machine-learned modules (preposition presence/choice, article presence/choice, auxiliary presence/choice, and gerund/infinitive confusion) all utilize a core set of features. Similar to work in contextual spelling correction (Golding & Roth, 1999), we take into account lexical and part-of-speech information from the context surrounding a potential error. The data we use for training are obtained from five different domains (see Table 1). Ideally, these training domains should closely resemble the kind of user input that one expects. However, in the absence of both an appropriate definition of targeted user text genre and large corpora reflecting typical well formed user input, we instead use widely available data sources that are somewhat diverse.

Table 1
Training Data for Machine-Learned Modules

Domain	Number of sentences
Encarta encyclopedia	487,281
Reuters newswire	567,394
United Nations proceedings	500,000
Europarl (European parliament proceedings)	500,000
Web scraped, using an algorithm similar to STRAND (Resnik & Smith, 2003)	500,000
Total	2,554,675

The feature extraction component is illustrated in Figure 2 below. Using the article presence/choice module as an example, for each sentence in the training corpus and its associated sequence of part-of-speech tags, we determine with a simple heuristic whether a given position in the sentence is a potential location for an article. Potential locations for articles are defined as left edges of a noun phrase which, in turn, is identified based on part-of-speech tags. For each such location, four tokens to the left and four tokens to the right, and six part-of-speech tags to the left and to the right are extracted as individual features. (For ease of illustration, the contextual window is set to three in Figure 2.) Each feature consists of a label (part-of-speech tag or token) followed by an indication of its position relative to the potential article position. "PRON_-2", for example, indicates that two tokens to the left of the potential article position there is a "PRON" part-of-speech tag. The potential location of an article (in this case the definite article *the*) in the sentence "*Most of the time, this works*" generates the following features:

part-of-speech context to the left:	PRON_-2, PREP_-1
part-of-speech context to the right:	NOUN_+1, COMMA_+2, PRON_+3
lexical context to the left:	Most_-2, of_-1
lexical context to the right:	time_+1, ,_+2, this_+3

In addition to this standard set of contextual features, we have added a few "custom" features that are designed to focus on salient properties of the context:

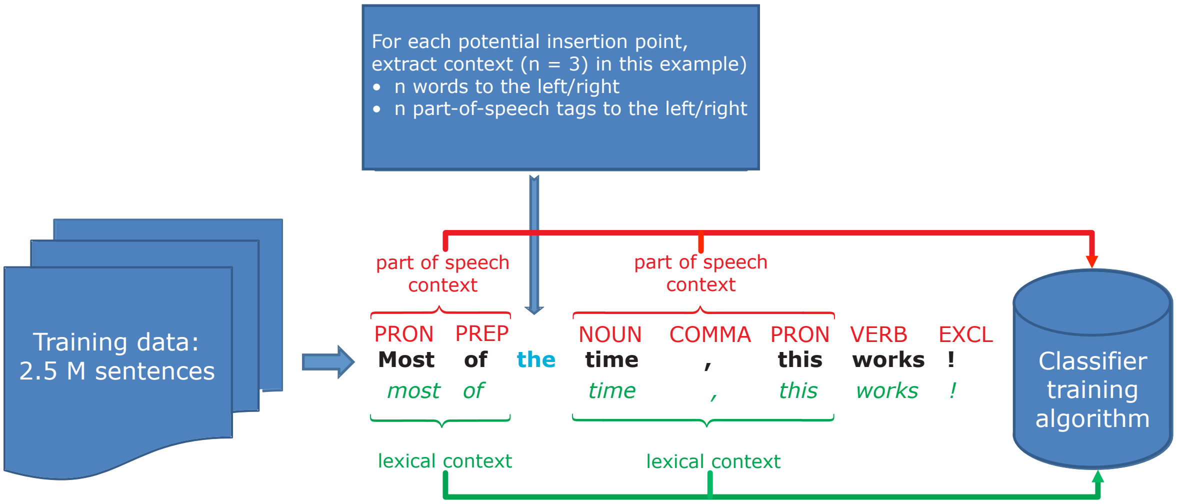
1. presence of capitalized tokens to the left or right,
2. presence of acronyms (tokens in all upper case),
3. mass noun/count noun status of the head of the noun phrase,
4. head of noun phrase, and
5. head of verb phrase.

These custom features were motivated by manual error analysis of cases where the classifier predicted a wrong preposition or article choice despite clear evidence to the contrary in the context. The capitalization-based features were introduced when we analyzed performance on email text that was rife with acronyms and capitalized out-of-vocabulary terms. Mass/count status and NP/VP heads allowed a more direct focus on subcategorization properties compared to just using distance-based features. The head of a verb phrase may be adjacent to a subcategorized token, but it could also be separated from that token by adverbs and other material. For each custom feature, we verified on a development set of nonnative text that precision would increase with little or no loss in recall.

Each feature vector generated from a potential article location in a sentence is then annotated with the information we are interested in predicting: Is there an article present in this potential location? If so, is it the definite or indefinite article?

The training set consists of all the feature vectors that are generated from the entire training corpus. In order to reduce the very large feature space, features that occur fewer than 10 times are eliminated. In a second pass of feature reduction, the log-likelihood ratio (Dunning, 1993) of each feature with respect to the article prediction task is determined, and only the top 75,000 features are retained.

Figure 2
Feature Extraction and Training for Machine-Learned Modules (For ease of illustration, the context window is set to three tokens to the left and right.)


For each potential insertion point, extract context (n = 3) in this example)
• n words to the left/right
• n part-of-speech tags to the left/right
Training data: 2.5 M sentences
part of speech context
part of speech context
PRON PREP NOUN COMMA PRON VERB EXCL
Most of the time , this works !
most of time , this works !
lexical context
lexical context
Classifier training algorithm

For article and preposition presence/choice, we chose to train two different classifiers: one predicts whether an article/preposition should be present in a given position, the second classifier predicts the choice of article/preposition, given that its presence has been determined. Choice of prepositions is limited to a set of prepositions that are both frequent and figure prominently in nonnative errors: *of*, *in*, *for*, *to*, *by*, *with*, *at*, *on*, *from*, *as*, *about*, and *since*. The decision to separate preposition and article classification into two stages is not based on any theoretical consideration. However, having a two-stage classification approach where presence/absence is determined before choice allows finer grained application of thresholds to each of the classification tasks. Finally, the presence/absence and choice classifiers are trained using a maximum entropy classifier.

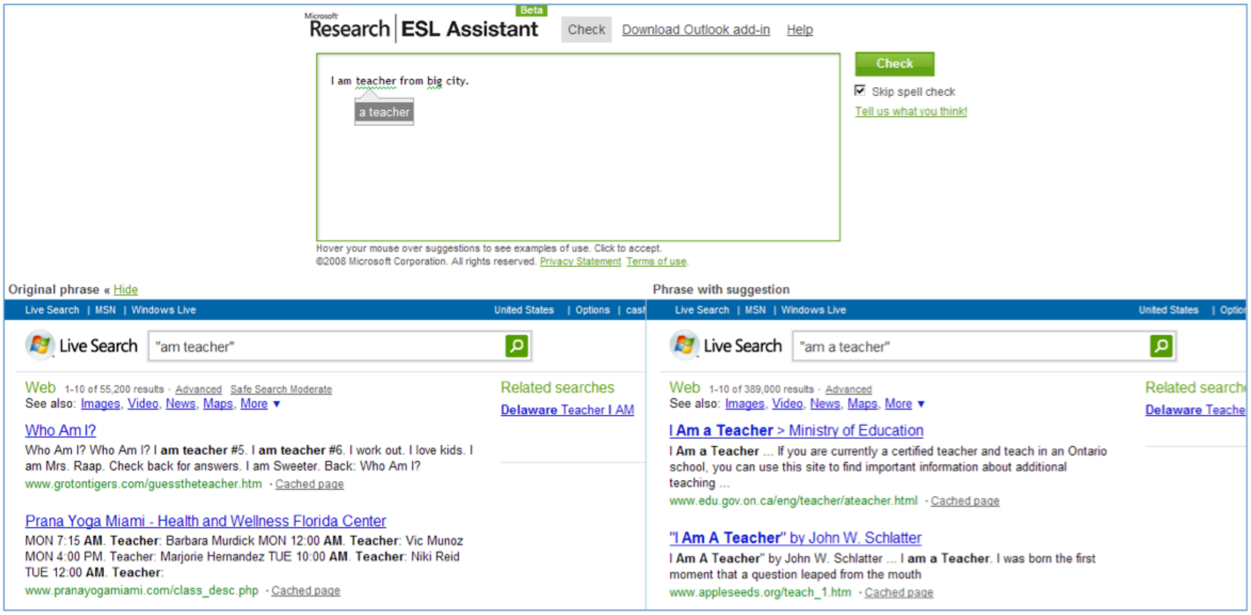
Utilizing large scale resources such as the Gigaword language model would not be practical on a standard desktop computer. For this reason, and in order to provide the web-based usage examples, ESL Assistant is implemented as a web service. A plugin for Microsoft Office Outlook is also available for download. The plugin allows the user to submit email text to the web service to check the text.

THE ROLE OF WEB-BASED EXAMPLES

There is plenty of anecdotal evidence that nonnative speakers use web search engines to find and verify usage of English expressions. The number of returned search results can serve as a proxy to identify "correctness." If, for example, a nonnative speaker is confused about whether to use *on the other hand* or *in the other hand*, a quick string search using a major web search engine for both expressions will yield roughly 60 million hits for *on the other hand* versus 250,000 hits for *in the other hand*. Furthermore, inspection of the results will quickly confirm that *on the other hand* is indeed the idiomatic expression, whereas *in the other hand*—while being perfectly well formed—is a literal expression. Previous research (Yi, Gao, & Dolan, 2008; Hermet, Désilets, & Szpakowicz, 2008) has shown that it is possible to use comparative web counts directly for error detection and correction, although the processing cost to issue multiple search queries for each potential error location and each potential correction is currently too high to make this approach practical. Searching for usage examples, however, only requires two search queries for each error that has been detected by the system: one query for the original input and the other for the suggested correction.

The ESL Assistant incorporates the search for usage examples as illustrated in Figure 3.

Figure 3
Side-by-Side Search Results for Original String and Suggested Correction



I am teacher from big city.

a teacher

Hover your mouse over suggestions to see examples of use. Click to accept.
©2008 Microsoft Corporation. All rights reserved. Privacy Statement Terms of use.

"am teacher"

1-10 of 55,200 results

Who Am I?
Who Am I? Who Am I? I **am teacher** #5. I **am teacher** #6. I work out. I love kids. I am Mrs. Raap. Check back for answers. I am Sweeter. Back: Who Am I?
www.grotontigers.com/guesstheteacher.htm · Cached page

Prana Yoga Miami - Health and Wellness Florida Center
MON 7:15 **AM**. **Teacher**: Barbara Murdick MON 12:00 **AM**. **Teacher**: Vic Munoz MON 4:00 PM. Teacher: Marjorie Hernandez TUE 10:00 **AM**. **Teacher**: Niki Reid TUE 12:00 **AM**. **Teacher**:
www.pranayogamiami.com/class_desc.php · Cached page

"am a teacher"

1-10 of 389,000 results

I **Am a Teacher** > Ministry of Education
I **Am a Teacher** ... If you are currently a certified teacher and teach in an Ontario school, you can use this site to find important information about additional teaching ...
www.edu.gov.on.ca/eng/teacher/ateacher.html · Cached page

"I **Am A Teacher**" by John W. Schlatter
I **Am A Teacher**" by John W. Schlatter ... I **am a Teacher**. I was born the first moment that a question leaped from the mouth
www.appleseeds.org/teach_1.htm · Cached page

When a potential error is detected and a suggested correction is offered, hovering over the suggested correction will trigger a side-by-side string search for both the original string and the suggested correction. This enables the user to verify whether the suggested correction corresponds to the writer's intent. Note that this functionality is especially useful when the correct choice of original words versus suggested correction is determined by semantic and/or pragmatic factors rather than clear-cut grammatical ones. In many cases, for example, the choice of a definite or indefinite determiner is conditioned by the larger discourse context, and examining this context may prove useful. In other cases, the pure frequency of the returned results for either option may be a good indicator for the "preferred" option. Preliminary user data indicate that for about 40% of the suggested corrections, users actually hover with their mouse over the suggested correction, triggering the parallel web search.

EVALUATION

In order to evaluate system performance, we use two evaluation strategies. First, individual modules can be tested on native text. For example, the prediction of preposition or article presence/choice should ideally be very close to the actual usage in native text. This strategy has the advantage that test data are plentiful and readily available and that testing can be fully automated. Based on the assumption that preposition and article usage in the native text represents the correct usage,[1] we can count an error each time the system predicts a preposition or article choice that is different from the observed choice, and we can count a correct prediction each time the predicted and actual choice are the same. This approach allows us to calculate precision and recall numbers without manual annotation. It does not, however, give us any sense of system performance on actual nonnative writing. For the latter purpose, we make use of the second approach—that of manual annotation—and evaluate performance on three domains of nonnative writing.

Evaluation on Native Text

In this section we report results for the two major machine-learned modules, article and preposition presence/choice on native text. For the evaluation, using the same mix of data that was used to train the classifiers, we split the data shown in Table 1 randomly into 70% training and 30% testing. We then trained our classifiers on the training set and tested the classifier performance on the held-out test set. The results for the article and preposition classifiers are shown in Tables 2-7. Both sets of classifiers achieve state-of-the-art performance compared to results reported in the literature: Turner and Charniak (2007), Han et al. (2004, 2006), Lee (2004), Minnen et al. (2000), and Knight and Chander (1994) for articles and Chodorow et al. (2007), Tetreault and Chodorow (2008a), and De Felice and Pulman (2007) for prepositions.

Table 2
Accuracy of the Article Classifiers

Classifier	Article presence/absence	Article choice	Combined
Accuracy	89.19	89.71	86.06

Table 3
Precision and Recall of the Article Presence/Absence Classifier

	Precision	Recall
Article presence	86.84	82.40
Article absence	90.40	93.00

Table 4
Precision and Recall of the Article Choice Classifier

	Precision	Recall
Definite	91.53	95.68
Indefinite	81.73	68.56

Table 5
Accuracy of the Preposition Classifiers

Classifier	Preposition presence/absence	Preposition choice	Combined
Accuracy	88.95	66.42	77.55

Table 6
Precision and Recall of the Preposition Presence/Absence Classifier

	Precision	Recall
Preposition presence	86.97	84.54
Preposition absence	90.18	91.81

Note that the prediction difficulty varies especially in the choice prediction for prepositions: the more “semantic” prepositions like *since* and *about* tend to be harder to predict correctly than the “grammatical” prepositions like *of* or *for* which are often more strictly conditioned by their environment and subcategorized for by a verb or noun.

Table 7
Precision and Recall of the Preposition Choice Classifier

	Precision	Recall
of	72.02	88.04
in	61.81	70.13
for	58.74	47.98
to	67.98	64.18
by	65.08	54.69
with	63.52	47.63
at	64.34	51.61
on	68.44	56.64
from	59.73	38.07
as	76.05	59.98
about	63.38	38.65
since	62.16	20.62
other	62.22	58.91

Evaluation on Nonnative Writing

While evaluation on native writing enables us to compare the ESL Assistant’s performance to that of other systems, we also need to complement this strategy by evaluating the system using nonnative writing in order to assess the system's "real-life" performance. However, this involves manual evaluation which is time consuming and therefore costly. In addition, human interrater agreement is known to be problematic, especially for articles and prepositions. While there would likely be very high interannotator agreement on some user errors, such as with overregularized verb inflection where the system rewrites *writed* as *wrote*, some error types are famously difficult to evaluate. Tetreault and Chodorow (2008b, p. 29) report that, for annotation of preposition errors, depending upon who is evaluating the errors, "using a single rater as a gold standard, there is the potential to over- or under-estimate precision by as much as 10%." On the other hand, evaluating the system on nonnative data is far more useful for system development than the relatively artificial task of predicting the articles and prepositions in native writing.

For the purposes of system development, we use a single annotator to evaluate the nonnative data. While the absolute numbers may not be as reliable for some modules as for others, the relative numbers for different domains and in improved or degraded performance are used to inform ESL Assistant's development.

To evaluate system performance on nonnative writing, we use data from three domains:[2]

1. The 1-million-word Chinese Learner's of English (CLEC) corpus (Gui & Yang, 2001, 2003)—a randomly selected subset of 10,000 sentences is used for blind testing.

2. A proprietary corpus of web-scraped nonnative English.
   These data were scraped from 489 personal web pages and blogs of nonnative speakers of English with Korean, Japanese, and Chinese language backgrounds. The data consists of 6,746 sentences, 1,000 of which were randomly selected for blind evaluation.

3. User data from a Microsoft-internal deployment of our service.
   These data were collected over 6 weeks from users who had installed an Outlook plugin allowing them to use our service to proofread their email. The final blind evaluation set contains 1,755 sentences.

The data we use for evaluation are in line with the design goal of producing a system that is geared towards native speakers of East Asian languages. An open question that we have not yet addressed is how useful this system is for users with different native language backgrounds. Relevant literature indicates that at least a subset of the common errors, such as preposition errors, are relatively independent of native language background (Dalgish, 1985; Bitchener, Young, & Cameron, 2005), so there is reason to believe that some of the error modules will prove useful outside of the area of East Asian native speakers.

Accuracy rates on nonnative text differ from those of native text because, instead of having a well formed sentence with a single decision to make, the nonnative sentences are often riddled with errors that confuse the classifier. In addition, the part-of-speech tagger that provides information to the classifier features is also likely to be affected by the noisy input.[3] Instead of the classifier being right or wrong, as when tested on copy-edited native writing, there is now a sizable 'neutral' space in which an error can be correctly identified but the correction is wrong, there is an error but the error type is misdiagnosed, there is a spelling error in the context, or when both the original and the suggested correction are acceptable. We therefore adopt a more detailed error categorization as shown in Table 8.

Table 8
Categorization of ESL Assistant Flags

Evaluation	Subevaluation	Description
Correct	Good flag	The correction fixes the problem.
Neutral	Both wrong	An error is correctly diagnosed, but the suggested rewrite does not correct the problem.
	Both OK	The original and the rewrite are both acceptable.
	Misdiagnosis	An error is to be found at or adjacent to the flagged location, but the system identifies a different kind of error (e.g., the article module identifies an error, but modifying the article is not the correct solution).
	Spell	A spelling error in the near context where Word 2008 did not suggest an appropriate choice.
Bad	False flag	The original is correct, but the rewrite is incorrect or inappropriate to the context.

The results for the machine-learned article and preposition modules across the three domains are shown in Figures 4 and 5. Domain is clearly a strong factor in classifier performance. However, the article module's performance is more stable across the domains than that of the preposition module. Whereas the article module generates a similar percentage of good flags across the three domains, the ratio of good preposition rewrites falls 20% from the CLEC corpus to the email corpus.

Figure 4
Performance of the Machine-Learned Article Module on Different Corpora

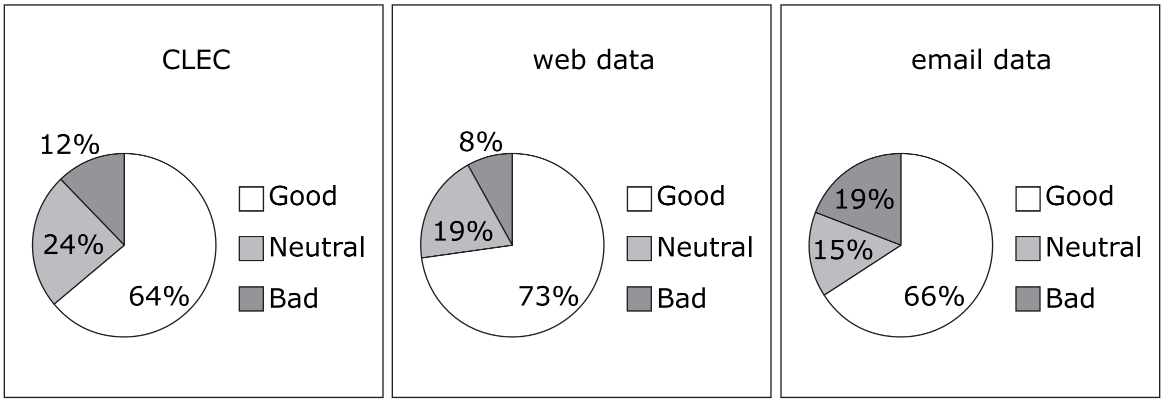
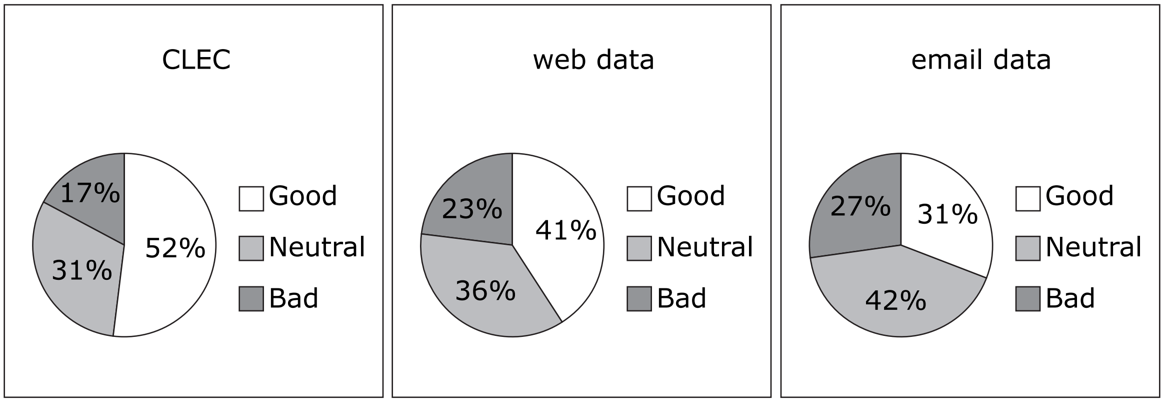


Figure 5
Performance of the Machine-Learned Preposition Module on Different Corpora



Neutral flags, which occur from 15% to 42% of the time, show that even when a suggestion does not improve the sentence, the modules often identify a problem spot. Although the module does not accurately diagnose or correct the error, a neutral flag can indicate the presence of another kind of error within the local context. For example, the string "*Let's compare them two*" generates a preposition flag suggesting that *with* be inserted between *them* and *two*, when in fact the real problem is an incorrect use of *them*.

A similar pattern emerges across all of the flags generated by ESL Assistant. Figures 6-9 below show the performance of all of the modules, grouped into verb-related, adjective-related, noun-related and preposition-related modules. These groups contain the following modules:

1. Verb-related modules
   We have already mentioned the machine-learned module that checks for the inclusion and choice of auxiliary verbs (*the situation *was/has changed much*) and the heuristic module that corrects overregularized verb inflection (*drived/drove*). Other verb-related modules cover a range of verb formation errors involving gerund/infinitive confusion (*is important *giving us* vs. *to give us*), perfect (*I have *studying/studied English*) and progressive tenses (*you are *hurry/hurrying*), passive sentences (*it will be *hold/held*), use of modal verbs (*they can *built/build a new house*), infinitives (**we want do* vs. *we want to do*), and confusion between a noun and verb (*I will *success/succeed*). Although the Word spell checker corrects many of these errors, it sometimes fails to identify them in a sentence that contains other errors because it cannot generate a parse—a situation often encountered in learner writing.

2. Adjective-related modules
   These modules address the word order of adjective sequences (*blue large bag* vs. *large blue bag*), adjective/noun confusions (**China/Chinese people*), and predicate adjective formation (*I am *interesting/interested in many things*).

3. Noun-related modules
   Along with the machine-learned articles module, there is another productive noun-related module that checks for noun number on plural forms of mass nouns (*save a lot of *labors/labor*) and on singular nouns (*not all *advertisement/advertisements are true*). Other modules address noun-of-noun constructions (**door of bus* versus *bus door*) and quantifiers on mass nouns (*The shop has *many business* vs. *a lot of business*).

4. Preposition-related modules
   In addition to the machine-learned preposition module, there is a heuristic preposition module that checks for phrasal verb constructions (**rely to a friend* vs. *rely on a friend*).

As a whole, the noun-related modules are relatively stable while the other modules are more sensitive to domain, as illustrated in Figures 6-9.

Figure 6
Performance of Verb-Related Modules on Different Corpora

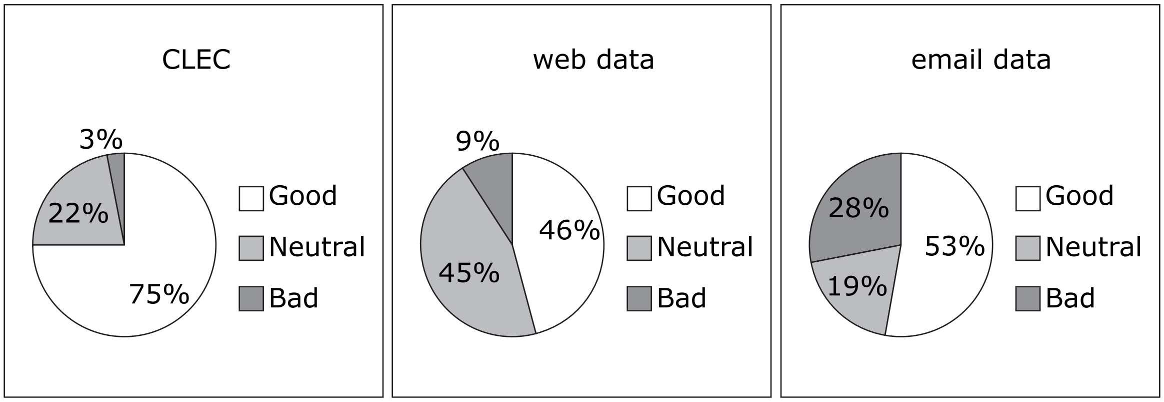


Figure 7
Performance of Adjective-Related Modules on Different Corpora

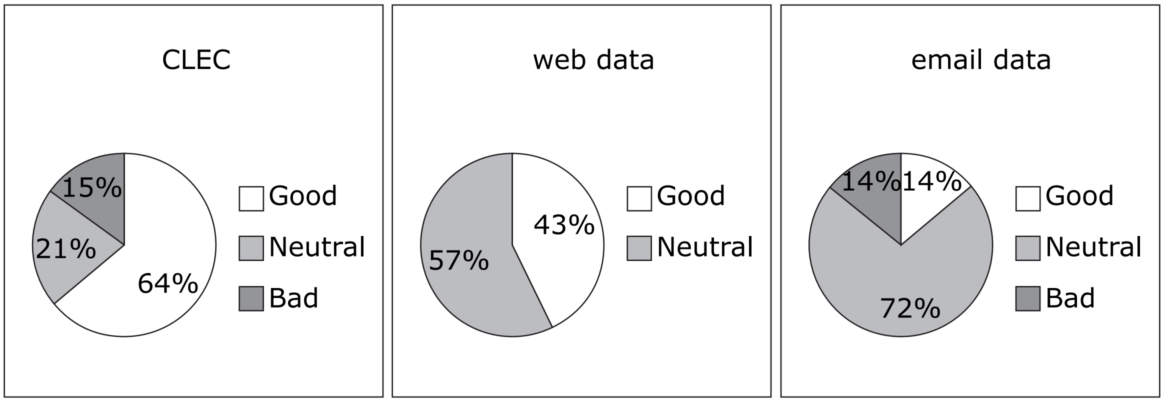


Figure 8
Performance of Noun-Related Modules on Different Corpora

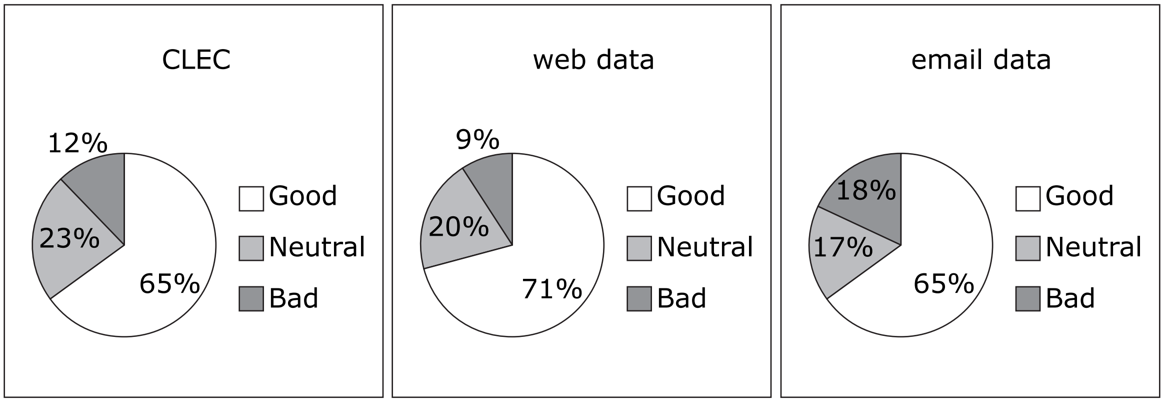
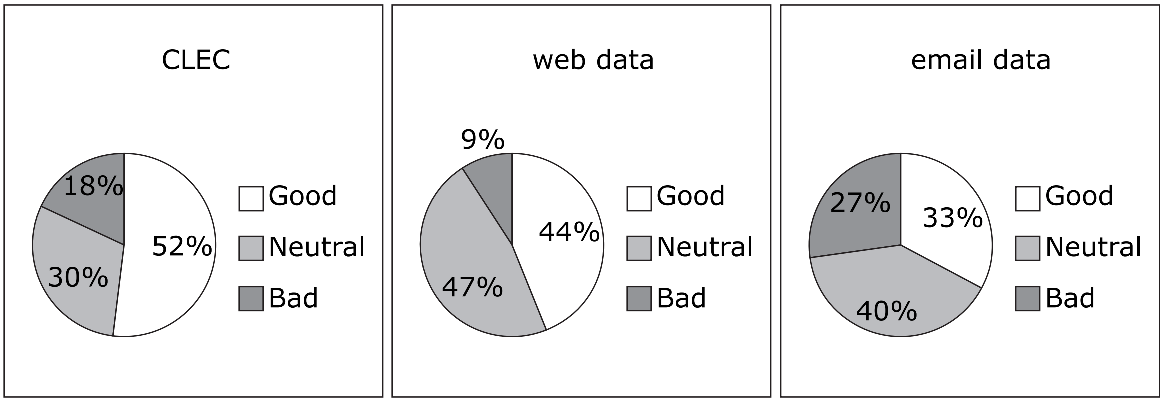


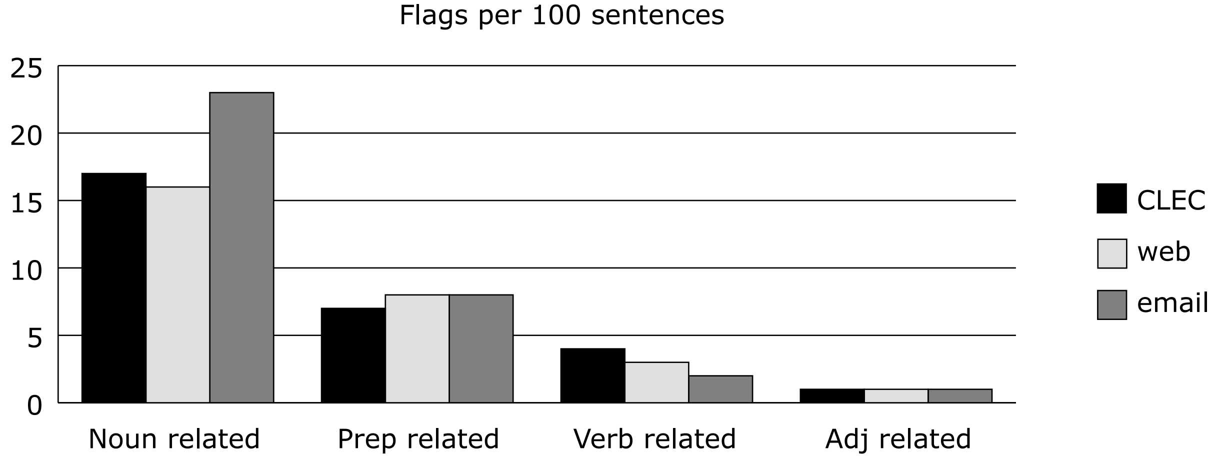
Figure 9
Performance of Preposition-Related Modules (both Machine-Learned and Heuristic) on Different Corpora



In the CLEC corpus, students are writing about themselves (e.g., their opinions, their interests, and their lives) in informal essay-style writing. Despite the fact that this corpus often has multiple errors within the same sentence, it appears to be the most amenable to corrections by ESL Assistant. Based on our error analysis we believe that the degraded performance on the email corpus that is observed across all modules is due to the relatively unique style of that corpus. The email corpus consists of business writing in the domain of a high-tech software company. Many of the false positives in this set are in the context of acronyms, product names (e.g., Excel, Vista), names for variables in computer programs (CheckOrgNameOverMatch), and the like. In addition there is computational sublanguage vocabulary such as "build 135" which ESL Assistant understandably wants to rewrite as "building 135." The high percentage of neutral flags for the adjective-related modules in the email corpus results from combinations of acronyms, abbreviations, and article errors that were present the 15 times that these modules fired. The web corpus contains a mixture of mostly colloquial writing: students describing themselves and their interests (as in CLEC) but also using a lot of acronyms and technical language (as in the email corpus). We are in the process of collecting a larger email corpus with which to continue system development.

The number of suggestions generated per 100 sentences is shown in Figure 10. Most of the suggestions generated by ESL Assistant are noun related (articles and noun number), followed by preposition-related suggestions. This follows the learner error ratios of the three most frequent JLE errors that are targeted by ESL Assistant: article errors comprise 27% of the JLE errors, preposition errors 10%, and noun number 8%. The pattern continues with the targeted verb-related errors comprising 10% of the learner errors and the targeted adjective-related errors 6%.

Figure 10
Number of Suggestions Generated by the Various Error Modules



To get some idea of ESL Assistant's recall, all of the errors in the 1,000 randomly selected sentences from the web page corpus were manually annotated for grammatical errors. ESL Assistant identified and corrected 37% of the 197 article errors that were identified in that corpus. Another noun-related module corrected 27% of the 77 errors involving noun number. The recall for the harder problem of preposition errors was lower, identifying and correcting 18% of the 151 preposition errors.

It is impossible to directly compare this system's performance with that of other systems. Reported results are based on different test corpora using different annotators and different annotation protocols. In addition, systems can make different kinds of suggestions. For example, Han et al. (2006) report 42% recall for detecting missing articles in TOEFL essays, but about half the time that system flags only that there is a missing article without identifying which article should be inserted. Along similar lines, Chodorow et al. (2007) report 21% and 26% recall (depending on the annotator) for preposition errors on a blind test set. Among other differences, the evaluation protocols differ in that when their classifier identifies a preposition that is in an ungrammatical context, the system is credited with correctly detecting a preposition error. We report these cases as being "other errors" and classify them in the neutral space. Given these differences, ESL Assistant's results compare favorably with those systems.

CONCLUSION

The ESL Assistant achieves state-of-the-art performance in detection and correction of the main types of errors made by nonnative speakers of English. This is accomplished by combining classification technology—rule-based components—and adding a large language model as a filter for suggested corrections. The results presented here are encouraging, and we hope that, even in its current state, the system will be useful for at least native speakers of East

Asian languages. Much work still remains to be done, however. Results vary considerably depending on the domain of user text, which points to the drawback of using generic native corpora for training. These native data are plentiful, but they fail to properly represent regularities of typical user input domains, such as email text or essay text. We believe that a next crucial step in the development of automatic proofing systems for nonnative speakers will be the collection of large amounts of real-life data.

NOTES

[1] This is, of course, an idealization since in some contexts multiple preposition and article choices may be equally acceptable, a fact that we ignore for our automatic evaluation. The reported numbers are lower bounds on the numbers one could obtain if every system prediction were manually checked.

[2] We did not include the NICT JLE data in the evaluation since our system is geared towards written language as opposed to spoken language.

[3] We do not have any evaluation of the tagger accuracy on nonnative corpora, but we are currently examining the performance of different taggers on "out of domain" nonnative text. We believe that whether a given tagger will be more or less robust to noise is a property that is orthogonal to the tagger's performance on the standard Penn Tree Bank data set. In other words, the tagger with the highest accuracy on the Penn Tree Bank may not be the best tagger when used out of domain.

REFERENCES

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, *14*(3), 191-205.

Bond, F., Ogura, K., & Ikehara, S. (1994). Countability and number in Japanese to English machine translation. In D. Coleman (Ed.), *Proceedings of the 15th Conference on Computational Linguistics* (pp. 32-38). Kyoto: Association for Computational Linguistics.

Chodorow, M., Tetreault, J. R., & Han, N.-R. (2007). Detection of grammatical errors involving prepositions. In F. Costello, J. Kelleher, & M. Volk (Eds.), *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions* (pp. 25-30). Prague: Association for Computational Linguistics.

Dalgish, G. M. (1985). Computer-assisted ESL research and courseware development. *Computers and Composition, 2*(4), 45-62.

De Felice, R., & Pulman, S. G. (2007). Automatically acquiring models of preposition use. In F. Costello, J. Kelleher, & M. Volk (Eds.), *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions* (pp. 45-50). Prague: Association for Computational Linguistics.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61-74.

Eeg-Olofsson, J., & Knutsson, O. (2003). Automatic grammar checking for second language learners—The use of prepositions. In *Proceedings of NoDaLiDa 2003*. Reykjavik, Iceland: Northern European Association for Language Technology.

Gao, J., Goodman, J., & Miao, J. (2001). The use of clustering techniques for language modeling—Application to Asian languages. *Computational Linguistics and Chinese Language Processing*, *6*(1), 27-60.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W. B., Belenko, D., & et al. (2008). Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing* (pp. 449-455). Hyderabad, India: Asian Federation of Natural Language Processing.

Golding, A. R., & Roth, D. (1999). A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, *34*(1), 107-130.

Gui, S., & Yang, H. (2001). *Computer analysis of Chinese learner English*. Paper presented at Hong Kong University of Science and Technology. Retrieved December 15, 2008, from http://lc.ust.hk/~centre/conf2001/keynote/subsect4/yang.pdf

Gui, S., & Yang, H. (Eds.). (2003). *Zhongguo Xuexizhe Yingyu Yuliaohu* [Chinese learner English corpus]. Shanghai: Shanghai Waiyu Jiaoyu Chubanshe.

Han, N.-R., Chodorow, M., & Leacock, C. (2004). Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 1625-1628). Lisbon: European Language Resources Association.

Han, N.-R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering, 12*(2), 115-129.

Heidorn, G. E. (2000). Intelligent writing assistance. In R. Dale, H. Moisl, & H. Somers (Eds.), *A handbook of natural language processing: Techniques and applications for the processing of language as text* (pp. 181-207). New York: Marcel Dekker.

Heine, J. E. (1998). Definiteness predictions for Japanese noun phrases. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 519-525). Montreal: Association for Computational Linguistics.

Hermet, M., Désilets, A., & Szpakowicz, S. (2008). Using the web as a linguistic resource to automatically correct lexico-syntactic errors. In *Proceedings of the Sixth International Language Resources and Evaluation* (pp. 390-396). Marrakech, Morocco: European Language Resources Association.

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., & Isahara, H. (2003). Automatic error detection in the Japanese learners' English spoken data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 145-148). Sapporo: Association for Computational Linguistics.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, *12*(2), 119-125.

Izumi, E., Uchimoto, K., & Isahara, H. (2005). Error annotation for corpus of Japanese learner English. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora* (pp. 71-80). Jeju Island, Korea: Association for Computational Linguistics.

Knight, K., & Chander, I. (1994). Automatic postediting of documents. In K. S. H. Forbus (Ed.), *Proceedings of the 12th National Conference on Artificial Intelligence* (pp. 779-784). Seattle: Morgan Kaufmann.

Lee, J. (2004). Automatic article restoration. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 31-36). Boston: Association for Computational Linguistics.

Linguistic Data Consortium (LDC). (2003). *English gigaword*. Available at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05

Minnen, G., Bond, F., & Copestake, A. (2000). Memory-based learning for article generation. In C. Cardie, W. Daelemans, C. Nédellec, & E. T. K. Sang (Eds.), *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop* (pp. 43-48). Lisbon: Association for Computational Linguistics.

Murata, M., & Nagao, M. (1993). Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation* (pp. 218-225). Kyoto: Kyoto International Community House.

Nagata, R., Wakana, T., Masui, F., Kawai, A., & Isu, N. (2005). Detecting article errors based on the mass count distinction. In R. Dale, W. Kam-Fie, J. Su, & O.Y. Kwong (Eds.), *Natural Language Processing-IJCNLP 2005, Second International Joint Conference Proceedings* (pp. 815-826). New York: Springer.

Nagata, R., Kawai, A., Morihiro, K., & Isu, N. (2006). A feedback-augmented method for detecting errors in the writing of learners of English. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 241-248). Sydney: Association for Computational Linguistics.

Nguyen, P., Gao, J., & Mahajan, M. (2007). *MSRLM: A scalable language modeling toolkit* (MSR-TR-2007-144). Redmond, WA: Microsoft.

Resnik, P., & Smith, N. (2003). The web as a parallel corpus. *Computational Linguistics, 29*(3), 349-380.

Tetreault, J. R., & Chodorow, M. (2008a). The ups and downs of prepositions. In *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 865-872). Manchester, UK: Association for Computational Linguistics.

Tetreault, J. R., & Chodorow, M. (2008b). Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Workshop on Human Judgments in Computational Linguistics, 22nd International Conference on Computational Linguistics* (pp 43-48). Manchester, UK: Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 252-259). Edmonton, Canada: Association for Computational Linguistics.

Turner, J., & Charniak, E. (2007). Language modeling for determiner selection. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*; *Companion Volume, Short Papers* (pp. 177-180). Rochester, NY: Association for Computational Linguistics.

Yi, X., Gao, J., & Dolan, W. B. (2008). A web-based English proofing system for English as a second language users. In *Proceedings of the Third International Joint Conference on Natural Language Processing* (pp. 619-624). Hyderabad, India: Asian Federation of Natural Language Processing.

AUTHORS' BIODATA

Michael Gamon has been a computational linguist in the Natural Language Processing Group, Microsoft Research since 1996. MA (linguistics) 1992, and Ph.D. (linguistics) 1996, University of Washington. His current research interests include automatic detection of language errors, modeling of linguistic properties, sentiment and subjectivity detection, and text analysis of social media.

Claudia Leacock is a consultant with the Butler Hill Group and is collaborating with Microsoft Research on the development of ESL Assistant. As a Distinguished Member of Technical Staff at Pearson Knowledge Technologies (2004-2007) and as a Principle Development Scientist at Educational Testing Service (1997-2004), she developed tools for both grammatical error detection/correction and automated assessment of short-answer content-based questions. As a member of the WordNet group at Princeton University's Cognitive Science Lab (1991-1997), her research focused on word sense identification. Dr. Leacock received a B.A. in English from NYU and a Ph.D. in Linguistics from the City University of New York, Graduate Center and was a post-doctoral fellow at IBM, T.J. Watson Research Center.

Chris Brockett has been a computational linguist in the Natural Language Processing Group, Microsoft Research since 1996. He has a B.A. in Asian Languages, Auckland, 1973; an M.A. in Japanese Literature, Waseda, 1979; and a Ph.D. in Linguistics, Cornell, 1991. Chris Brockett is lead author of *A Communicative Framework for Introductory Japanese Language Curricula* (University of Hawai'i Press, 2000). His current research interests include semantic networks and lexical similarity, paraphrase acquisition and generation, and writing assistance applications.

William B. Dolan has been Principal Researcher and Manager of the Natural Language Processing Group, Microsoft Research since 1992. He has an M.A. (Linguistics) 1987 and a Ph.D. (linguistics) 1994, University of California, Los Angeles. His current research interests include paraphrase recognition and generation, semantic networks and lexical similarity, and writing assistance applications.

Jianfeng Gao has been a researcher in the Natural Language Processing Group, Microsoft Research since 2006. From 2005 to 2006, he was a software developer in the Natural Interactive Services Division at Microsoft. From 1999 to 2005, he was a researcher in the Natural Language Computing Group, Microsoft Research Asia. His current research interests include web search and mining, information retrieval, natural language processing, and statistical machine learning.

Dmitriy Belenko has been a Research Software Development Engineer in the Natural Language Processing Group, Microsoft Research since 2007. He has a M.Sc. (CS/EE) from the Moscow State Technical University, Moscow, Russia. His current research interests include machine learning and its applications in linguistics, information retrieval, and the combination of the two.

Alexandre Klementiev is a Ph.D. student at the University of Illinois at Urbana-Champaign in the Department of Computer Science working with Prof. Dan Roth. His research interests are on the intersection of machine learning and natural language processing.

AUTHORS' ADDRESSES

Michael Gamon
One Microsoft Way
Redmond, WA 98052
Phone: 425 703 2976
Fax: 425 936 7329
Email: mgamon@microsoft.com

Claudia Leacock
100 Bleecker Street, 27A
New York, NY 10012
Phone and fax: 212 674 1989
Email: claudia.leacock@gmail.com

Chris Brockett
One Microsoft Way
Redmond, WA 98052
Phone: 425 703 2976
Fax: 425 936 7329
Email: chrisbkt@microsoft.com

William B. Dolan
One Microsoft Way
Redmond, WA 98052
Phone: 425 706-3709
Fax: 425 936 7329
Email: billdol@microsoft.com

Jianfeng Gao
One Microsoft Way
Redmond, WA 98052
Phone: 425 705 1479
Fax: 425 936 7329
Email: jfgao@microsoft.com

Dmitriy Belenko
One Microsoft Way
Redmond, WA 98052
Phone: 425 707 6994
Fax: 425 936 7329
Email: dmitryb@microsoft.com

Alexandre Klementiev
201 N. Goodwin Ave.,
Urbana, IL 61801
Phone: 217 333 2584
Email: klementi@uiuc.edu