The Illinois-Columbia System in the CoNLL-2014 Shared Task

Alla Rozovskaya[1] Kai-Wei Chang[2] Mark Sammons[2] Dan Roth[2] Nizar Habash[1]

[1]Center for Computational Learning Systems, Columbia University
{alla,habash}@ccls.columbia.edu
[2] Cognitive Computation Group, University of Illinois at Urbana-Champaign
{kchang10,mssammon,danr}@illinois.edu

Abstract

The CoNLL-2014 shared task is an extension of last year's shared task and focuses on correcting grammatical errors in essays written by non-native learners of English. In this paper, we describe the Illinois-Columbia system that participated in the shared task. Our system ranked second on the original annotations and first on the revised annotations.

The core of the system is based on the University of Illinois model that placed first in the CoNLL-2013 shared task. This baseline model has been improved and expanded for this year's competition in several respects. We describe our underlying approach, which relates to our previous work, and describe the novel aspects of the system in more detail.

1 Introduction

The topic of text correction has seen a lot of interest in the past several years, with a focus on correcting grammatical errors made by English as a Second Language (ESL) learners. ESL error correction is an important problem since most writers of English are not native English speakers. The increased interest in this topic can be seen not only from the number of papers published on the topic but also from the three competitions devoted to grammatical error correction for non-native writers that have recently taken place: HOO-2011 (Dale and Kilgarriff, 2011), HOO-2012 (Dale et al., 2012), and the CoNLL-2013 shared task (Ng et al., 2013).

In all three shared tasks, the participating systems performed at a level that is considered extremely low compared to performance obtained in other areas of NLP: even the best systems attained F1 scores in the range of 20-30 points.

The key reason that text correction is a difficult task is that even for non-native English speakers, writing accuracy is very high, as *errors are very sparse*. Even for some of the most common types of errors, such as article and preposition usage, the majority of the words in these categories (over 90%) are used correctly. For instance, in the CoNLL training data, only 2% of prepositions are incorrectly used. Because errors are so sparse, it is more difficult for a system to identify a mistake accurately and without introducing many false alarms.

The CoNLL-2014 shared task (Ng et al., 2014) is an extension of the CoNLL-2013 shared task (Ng et al., 2013). Both competitions make use of essays written by ESL learners at the National University of Singapore. However, while the first one focused on five kinds of mistakes that are commonly made by ESL writers – article, preposition, noun number, verb agreement, and verb form – this year's competition covers all errors occurring in the data. Errors outside the target group were present in the task corpora last year as well, but were not evaluated.

Our system extends the one developed by the University of Illinois (Rozovskaya et al., 2013) that placed first in the CoNLL-2013 competition. For this year's shared task, the system has been extended and improved in several respects: we extended the set of errors addressed by the system, developed a general approach for improving the error-specific models, and added a joint inference component to address interaction among errors. See Rozovskaya and Roth (2013) for more detail.

We briefly discuss the task (Section 2) and give an overview of the baseline Illinois system (Section 3). Section 4 presents the novel aspects of the system. In Section 5, we evaluate the complete system on the development data and show the results obtained on test. We offer error analysis and a brief discussion in Section 6. Section 7 concludes.

Error type	Rel. freq.	Examples
Article (_ArtOrDet_)	14.98%	*∅/_The_ government should help encourage *the/∅ breakthroughs as well as *a/∅ complete medication system .
Wrong collocation (_Wci_)	11.94%	Some people started to *_think/wonder_ if electronic products can replace human beings for better performances .
Local redundancy (_Rloc-_)	10.52%	Some solutions *{_as examples_}/∅ would be to design plants/fertilizers that give higher yield ...
Noun number (_Nn_)	8.49%	There are many reports around the internet and on newspaper stating that some users ' *_iPhone/iPhones_ exploded .
Verb tense (_Vt_)	7.21%	Through the thousands of years , most Chinese scholars *_are_/{_have been_} greatly affected by Confucianism .
Orthography/punctuation (_Mec_)	6.88%	Even British Prime Minister , Gordon Brown *∅/, has urged that all cars in *_britain/Britain_ to be green by 2020 .
Preposition (_Prep_)	5.43%	I do not agree *_on/with_ this argument that surveillance technology should not be used to track people .
Word form (_Wform_)	4.87%	On the other hand , the application of surveillance technology serves as a warning to the *_murders/murderers_ and they might not commit more murder .
Subject-verb agreement (_SVA_)	3.44%	However , tracking people *_are/is_ difficult and different from tracking goods .
Verb form (_Vform_)	3.25%	Travelers survive in desert thanks to GPS *_guide/guiding_ them .
Tone (_Wtone_)	1.29%	Hence , as technology especially in the medical field continues to get developed and updated , people {_do n't_}/{_do not_} risk their lives anymore .

Table 1: **Example errors.** In the parentheses, the error codes used in the shared task are shown. Note that only the errors exemplifying the relevant phenomena are marked in the table; the sentences may contain other mistakes. Errors marked as verb form include multiple grammatical phenomena that may characterize verbs. Our system addresses all of the error types except “Wrong Collocation” and “Local Redundancy”.

2 Task Description

Both the training and the test data of the CoNLL-2014 shared task consist of essays written by students at the National University of Singapore. The training data contains 1.2 million words from the NUCLE corpus (Dahlmeier et al., 2013) corrected by English teachers, and an additional set of about 30,000 words that was released last year as a test set for the CoNLL-2013 shared task. We use last year’s test data as a _development_ set; the results in the subsequent sections are reported on this subset.

The CoNLL corpus error tagset includes 28 error categories. Table 1 illustrates the most common error categories in the training data; errors are marked with an asterisk, and ∅ denotes a missing word. Our system targets all of these, with the exception of collocation and local redundancy errors. Among the less commonly occurring error types, our system addresses tone (style) errors; these are illustrated in the table.

It should be noted that the proportion of erroneous instances is several times higher in the development data than in the training data for all of the error categories. For example, while only 2.4% of noun phrases in the training data have determiner errors, in the development data 10% of noun phrases have determiner errors.

<table>
<tr><td colspan="2">“Hence, the environmental *<b>factor/factors</b> also *<b>contributes/contribute</b> to various difficulties, *<b>included/including</b> problems in nuclear technology.”</td></tr>
<tr><th>Error type</th><th>Confusion set</th></tr>
<tr><td>Noun number</td><td>{factor, factors}</td></tr>
<tr><td>Verb Agreement</td><td>{contribute, contributes}</td></tr>
<tr><td>Verb Form</td><td>{included, including, includes, include}</td></tr>
</table>

Table 2: **Sample confusion sets for noun number, verb agreement, and verb form.**

3 The Baseline System

In this section, we briefly describe the University of Illinois system (henceforth *Illinois*; in the overview paper of the shared task the system is referred to as UI) that achieved the best result in the CoNLL-2013 shared task and which we use as our baseline model. For a complete description, we refer the reader to Rozovskaya et al. (2013).

The Illinois system implements five independently-trained machine-learning classifiers that follow the popular approach to ESL error correction borrowed from the context-sensitive spelling correction task (Golding and Roth, 1999; Carlson et al., 2001). A *confusion set* is defined as a list of confusable words. Each occurrence of a confusable word in text is represented as a vector of features derived from a context window around the target. The problem is cast as a multi-class classification task and a classifier is trained on native or learner data. At prediction time, the model selects the most likely candidate from the confusion set.

The confusion set for prepositions includes the top 12 most frequent English prepositions (this year, we extend the confusion set and also target extraneous preposition usage). The article confusion set is as follows: $\{a, the, \varnothing\}$.[1] The confusion sets for *noun*, *agreement*, and *form* modules depend on the target word and include its morphological variants. Table 2 shows sample confusion sets for *noun*, *agreement*, and *form* errors.

Each classifier takes as input the corpus documents preprocessed with a part-of-speech tagger[2] and shallow parser[3] (Punyakanok and Roth, 2001). The other system components use the preprocessing tools only as part of candidate generation (e.g., to identify all nouns in the data for the noun classifier).

The choice of learning algorithm for each classifier is motivated by earlier findings showing that discriminative classifiers outperform other machine-learning methods on error correction tasks (Rozovskaya and Roth, 2011). Thus, the classifiers trained on the learner data make use of a discriminative model. Because the Google corpus does not contain complete sentences but only n-gram counts of length up to five, training a discriminative model is not desirable, and we thus use NB (details in Rozovskaya and Roth (2011)).

The *article* classifier is a discriminative model that draws on the state-of-the-art approach described in Rozovskaya et al. (2012). The model makes use of the Averaged Perceptron (AP) algorithm (Freund and Schapire, 1996) and is trained on the training data of the shared task with rich features. The article module uses the POS and chunker output to generate some of its features and candidates (likely contexts for missing articles).

The original word choice (the source article) used by the writer is also used as a feature. Since the errors are sparse, this feature causes the model to abstain from flagging mistakes, resulting in low recall. To avoid this problem, we adopt the approach proposed in Rozovskaya et al. (2012), the *error inflation* method, and add artificial article errors to the training data based on the error distribution on the training set. This method prevents the source feature from dominating the context features, and improves the recall of the system.

The other classifiers in the baseline system – noun number, verb agreement, verb form, and preposition – are trained on native English data, the Google Web 1T 5-gram corpus (henceforth, Google, (Brants and Franz, 2006)) with the Naïve Bayes (NB) algorithm. All models use word n-gram features derived from the 4-word window around the target word. In the *preposition* model, priors for preposition preferences are learned from the shared task training data (Rozovskaya and Roth, 2011).

The modules targeting *verb agreement* and

[1] $\varnothing$ denotes noun-phrase-initial contexts where an article is likely to have been omitted. The variants “a” and “an” are conflated and are restored later.

[2] `http://cogcomp.cs.illinois.edu/page/software_view/POS`

[3] `http://cogcomp.cs.illinois.edu/page/software_view/Chunker`

verb form mistakes draw on the linguistically-motivated approach to correcting verb errors proposed in Rozovskaya et. al (2014).

4 The CoNLL-2014 System

The system in the CoNLL-2014 shared task is improved in three ways: 1) Additional error-specific classifiers: word form, orthography/punctuation, and style; 2) Model combination; and 3) Joint inference to address interacting errors. Table 3 summarizes the Illinois and the Illinois-Columbia systems.

4.1 Targeting Additional Errors

The Illinois-Columbia system implements several new classifiers to address word form, orthography and punctuation, and style errors (Table 1).

4.1.1 Word Form Errors

Word form (*Wform*) errors are grammatical errors that involve confusing words that share a base form but differ in derivational morphology, e.g. "use" and "usage" (see also Table 1). Confusion sets for word form errors thus should include words that differ derivationally but share the same base form. In contrast to verb form errors where confusion sets specify all possible inflectional forms for a given verb, here, the associated parts-of-speech may vary more widely. An example of a confusion set is {technique, technical, technology, technological}.

Because word form errors encompass a wide range of misuse, one approach is to consider every word as an error candidate. We follow a more conservative method and only attempt to correct those words that occurred in the training data and were tagged as word form errors (we cleaned up that list by removing noisy annotations).

A further challenge in addressing word form errors is generating confusion sets. We found that about 45% of corrections for word form errors in the development data are covered by the confusion sets from the training data for the same word. We thus derive the confusion sets using the training data. Specifically, for every source word that is tagged as a word form error in the training data, the confusion set includes all labels to which that word is mapped in the training data. In addition, plural and singular forms are added for all words tagged as nouns, and inflectional forms are added for words tagged as verbs. For more detail on correcting verb errors, we refer the reader to Rozovskaya et al. (2014).

4.1.2 Orthography and Punctuation Errors

The *Mec* error category includes errors in spelling, context-sensitive spelling, capitalization, and punctuation. Our system addresses punctuation errors and capitalization errors.

To correct capitalization errors, we collected words that are always capitalized in the training and development data when not occurring sentence-initially.

The punctuation classifier includes two modules: a learned component targets missing and extraneous comma usage and is an AP classifier trained on the learner data with error inflation. A second, pattern-based component, complements the AP model: it inserts missing commas by using a set of patterns that overwhelmingly prefer the usage of a comma, e.g. when a sentence starts with the word "hence". The patterns are learned automatically over the training data: specifically, using a sliding window of three words on each side, we compiled a list of word n-gram contexts that are strongly associated with the usage of a comma. This list is then used to insert missing commas in the test data.

4.1.3 Style Errors

The style (*Wtone*) errors marked in the corpus are diverse, and the annotations are often not consistent. We constructed a pattern-based system to deal with two types of *style* errors that are commonly annotated. The first type of *style* edit avoids using contractions of negated auxiliary verbs. For example, it changes "do n't" to "do not". We use a pattern-based classifier to identify such errors and replace the contractions. The second type of *style* edit encourages the use of a semi-colon to join two independent clauses when a conjunctive adverb is used. For example, it edits "[clause], however, [clause]" to "[clause]; however, [clause]". To identify such errors, we use a part-of-speech tagger to recognize conjunctive adverbs signifying independent clauses: if two clauses are joined by the pattern ", [conjunctive adverb],", we will replace it with "; [conjunctive adverb],".

4.2 Modules not Included in the Final System

In addition to the modules described above, we attempted to address two other common error categories: spelling errors and collocation errors. We

Illinois		
Classifiers	**Training data**	**Algorithm**
Article	Learner	AP with inflation
Preposition	Native	NB-priors
Noun number	Native	NB
Verb agreement	Native	NB
Verb form	Native	NB

Illinois-Columbia		
Classifiers	**Training data**	**Algorithm**
Article	Learner and native	AP with infl. (learner) and NB-priors (native)
Preposition	Learner and native	AP with infl. (learner) and NB-priors (native)
Noun number	Learner and native	AP with infl. (learner) and NB (native)
Verb agreement	Native	AP with infl. (learner) and NB (native)
Verb form	Native	NB-priors
Word form	Native	NB-priors
Orthography/punctuation	Learner	AP and pattern-based
Style	Learner	Pattern-based
Model combination	Section 4.3	
Global inference	Section 4.4	

Table 3: **The baseline (Illinois) system vs. the Illinois-Columbia system.** AP stands for Averaged Perceptron, and NB stands for the Naïve Bayes algorithm.

describe these below even though they were not included in the final system.

Regular spelling errors are noticeable but not very frequent, and a number are not marked in the corpus (for example, the word "dictronary" instead of "dictionary" is not tagged as an error). We used an open source package – "Jazzy"[4] – to attempt to automatically correct these errors to improve context signals for other modules. However, there are often multiple similar words that can be proposed as corrections, and Jazzy uses phonetic guidelines that sometimes lead to unintuitive proposals (such as "doctrinaire" for "dictronary"). It would be possible to extend the system with a filter on candidate answers that uses n-grams or some other context model to choose better candidates, but the relatively small number of such errors limits the potential impact of such a system.

Collocation errors are the second most common error category accounting for 11.94% of all errors in the training data (Table 1). We tried using the Illinois context-sensitive spelling system[5] to detect these errors, but this system requires predefined confusion sets to detect possible errors and to propose valid corrections. The coverage of the pre-existing confusion sets was poor – the system could potentially correct only 2.5% of collocation errors – and it is difficult to generate new confusion sets that generalize well, which requires a great deal of annotated training data. The system performance was relatively poor because it proposed many spurious corrections: we believe this is due to the relatively limited context it uses, which makes it particularly susceptible to making mistakes when there are multiple errors in close proximity.

4.3 Model Combination

Model combination is another key extension of the Illinois system.

In the Illinois-Columbia system, article, preposition, noun, and verb agreement errors are each addressed via a model that combines error predictions made by a classifier trained on the learner data with the AP algorithm and those made by the NB model trained on the Google corpus. The AP classifiers all make use of richer sets of features than the native-trained classifiers: the article, noun number, and preposition classifiers employ features that use POS information, while the verb agreement classifier also makes use of dependency features extracted using a parser (de Marneffe et al., 2008). For more detail on the features used in the agreement module, we refer the reader to

[4] http://jazzy.sourceforge.net

[5] http://cogcomp.cs.illinois.edu/cssc/

Rozovskaya et al. (2014). Finally, all of the AP models use the source word of the author as a feature and, similar to the article AP classifier (Section 3), implement the error inflation method. The combined model generates a union of corrections produced by the components.

We found that for every error type, the combined model is superior to each of the single classifiers, as it combines the advantages of both of the classifiers so that they complement one another. In particular, while each of the learner and native components have similar precision, since the predictions made differ, the recall of the combined model improves.

4.4 Joint Inference

One of the mistakes typical for Illinois system were inconsistent predictions. Inconsistent predictions occur when the classifiers address grammatical phenomena that interact at the sentence level, e.g. noun number and verb agreement. To address this problem, the Illinois-Columbia system makes use of global inference via an Integer Linear Programming formulation (Rozovskaya and Roth, 2013). Note that Rozovskaya and Roth (2013) also describe a joint learning model that performs better than the joint inference approach. However, the joint learning model is based on training a joint model on the Google corpus, and is not as strong as the individually-trained classifiers of the Illinois-Columbia system that combine predictions from two components – NB classifiers trained on the native data from the Google corpus and AP models trained on the learner data (Section 4.3).

5 Experimental Results

In Sections 3 and 4, we described the individual system components that address different types of errors. In this section, we show how the system improves when each component is added into the system. In this year's competition, systems are compared using F0.5 measure instead of F1. This is because in error correction good precision is more important than having a high recall, and the F0.5 reflects that by weighing precision twice as much as recall. System output is scored with the M2 scorer (Dahlmeier and Ng, 2012).

Table 4 reports performance results of each individual classifier. In the final system, the article, preposition, noun number, and verb agree-

Model	**P**	**R**	**F0.5**
Articles (AP)	38.97	8.85	23.19
Articles (NB-priors)	47.34	6.01	19.93
Articles (Comb.)	38.73	10.93	25.67
Prep. (AP)	34.00	0.5	2.35
Prep. (NB-priors)	33.33	0.79	3.61
Prep. (Comb.)	30.06	1.17	5.13
Noun number (NB)	44.74	5.48	18.39
Noun number (AP)	82.35	0.41	2.01
Noun number (Comb.)	45.02	5.57	18.63
Verb agr. (AP)	38.56	1.23	5.46
Verb agr. (NB)	63.41	0.76	3.64
Verb agr. (Comb.)	41.09	1.55	6.75
Verb form (NB-priors)	59.26	1.41	6.42
Word form (NB-priors)	57.54	3.02	12.48
Mec (AP; patterns)	48.48	0.47	2.26
Style (patterns)	84.62	0.64	3.13

Table 4: **Performance of classifiers targeting specific errors**.

Model	**P**	**R**	**F0.5**
The baseline (Illinois) system			
Articles	38.97	8.85	23.19
+Prepositions	39.24	9.35	23.93
+Noun number	42.13	14.83	30.79
+Subject-verb agr.	42.25	16.06	31.86
+Verb form	43.19	17.20	33.17
Model Combination			
+Model combination	42.72	20.19	34.92
Additional Classifiers			
+Word form	43.39	21.54	36.07
+Mec	43.70	22.04	36.52
+Style	44.22	21.54	37.09
Joint Inference			
+Joint Inference	44.28	22.57	37.13

Table 5: **Results on the development data**. The top part of the table shows the performance of the baseline (Illinois) system from last year.

P	**R**	**F0.5**
Scores based on the original annotations		
41.78	24.88	36.79
Scores based on the revised annotations		
52.44	29.89	45.57

Table 6: **Results on Test.**

ment classifiers use combined models, each consisting of a classifier trained on the learner data and a classifier trained on native data. We report performance of each such component separately and when they are combined. The results show that combining models boosts the performance of each classifier: for example, the performance of the article classifier improves by more than 2 F0.5 points. It should be noted that results are computed with respect to all errors present in the data. For this reason, recall is low.

Next, in Table 5, we show the contribution of the novel components over the baseline system on the development set. As described in Section 3, the baseline Illinois system consists of five individual components; their performance is shown in the top part of the table. Note that although for the development set we make use of last year's test set, these results are not comparable to the performance results reported in last year's competition that used the F1 measure. Overall, the baseline system achieves an F0.5 score of 33.17 on the development set.

Then, by applying the model combination technique introduced in Section 4.3, the performance is improved to 34.92. By adding modules to target three additional error types, the overall performance becomes 37.09. Finally, the joint inference technique (see Section 4.4) slightly improves the performance further. The final system achieves an F0.5 score of 37.13.

Table 6 shows the results on the test set provided by the organizers. As was done previously, the organizers also offered another set of annotations based on the combination of revised official annotations and accepted alternative annotations proposed by participants. Performance results on this set are also shown in Table 6.

6 Discussion and Error Analysis

Here, we present some interesting errors that our system makes on the development set and discuss our observations on the competition. We analyze both the false positive errors and those cases that are missed by our system.

6.1 Error Analysis

Stylistic preference *Surveillance* ***technology*** *such as RFID (radio-frequency identification) is one type of examples that has currently been implemented.*

Here, our system proposes a change to plural for the noun "technology". The gold standard solution instead proposes a large number of corrections throughout that work with the choice of the singular "technology". However, using the plural "technologies" as proposed by the Illinois-Columbia system is quite acceptable, and a comparable number of corrections would make the rest of the sentence compatible. Note also that the gold standard proposes the use of commas around the phrase "such as RFID (radio-frequency identification)", which could also be omitted based on stylistic considerations alone.

Word choice *The high accuracy in utilizing surveillance technology eliminates the* *amount/number *of disagreements among people.*

The use of "amount" versus "number" depends on the noun to which the term attaches. This could conceivably be achieved by using a rule and word list, but many such rules would be needed and each would have relatively low coverage. Our system does not detect this error.

Presence of multiple errors *Not only the details of location will be provided, but also may lead to find out the root of this kind of children trading agency and it helps to prevent more this kind of tragedy to happen* ***on*** *any family.*

The writer has made numerous errors in this sentence. To determine the correct preposition in the marked location requires at least the preceding verb phrase to be corrected to "from happening"; the extraneous "more" after "prevent" in turn makes the verb phrase correction more unlikely as it perturbs the contextual clues that a system might learn to make that correction. Our system proposes a different preposition – "in" – that is better than the original in the local context, but which is not correct in the wider context.

Locally coherent, globally incorrect *People's lives become* ***from*** *increasingly convenient to almost luxury, thanks to the implementation of increasingly technology available for the Man's life.*

In this example, the system proposes to delete the preposition "from". This correctiom improves the local coherency of the sentence. However, the resulting construction is not consistent with "to almost luxury", suggesting a more complex correction (changing the word "become" to "are going").

Cascading NLP errors *In this, I mean that we can input this device* ***implant*** *into an animal or birds species, for us to track their movements and actions relating to our human research that can bring us to a new regime.*

The word "implant" in the example sentence has been identified as a verb by the system and not a noun due to the unusual use as part of the phrase "device implant". As a result, the system incorrectly proposes the verb form correction "implanted".

6.2 Discussion

The error analysis suggests that there are three significant challenges to developing a better grammar correction system for the CoNLL-2014 shared task: identifying candidate errors; modeling the context of possible errors widely enough to capture long-distance cues where necessary; and modeling stylistic preferences involving word choice, selection of plural or singular, standards for punctuation, use of a definite or indefinite article (or no article at all), and so on. For ESL writers, the tendency for multiple errors to be made in close proximity means that global decisions must be made about sets of possible mistakes, and a system must therefore have a quite sophisticated abstract model to generate the basis for consistent sets of corrections to be proposed.

7 Conclusion

We have described our system that participated in the shared task on grammatical error correction. The system builds on the elements of the Illinois system that participated in last year's shared task. We extended and improved the Illinois system in three key dimensions, which we presented and evaluated in this paper. We have also presented error analysis of the system output and discussed possible directions for future work.

Acknowledgments

This material is based on research sponsored by DARPA under agreement number FA8750-13-2-0008. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. This research is also supported by a grant from the U.S. Department of Education and by the DARPA Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-018. The first and last authors were partially funded by grant NPRP-4-1058-1-168 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

T. Brants and A. Franz. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, PA.

A. J. Carlson, J. Rosen, and D. Roth. 2001. Scaling up context sensitive text correction. In *IAAI*.

D. Dahlmeier and H.T. Ng. 2012. Better evaluation for grammatical error correction. In *NAACL*, pages 568–572, Montréal, Canada, June. Association for Computational Linguistics.

D. Dahlmeier, H.T. Ng, and S.M. Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proc. of the NAACL HLT 2013 Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, June. Association for Computational Linguistics.

R. Dale and A. Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*.

R. Dale, I. Anisimoff, and G. Narroway. 2012. A report on the preposition and determiner error correction shared task. In *Proc. of the NAACL HLT 2012 Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, Montreal, Canada, June. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *ACL*.

Yoav Freund and Robert E. Schapire. 1996. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*.

A. R. Golding and D. Roth. 1999. A Winnow based approach to context-sensitive spelling correction. *Machine Learning*.

H.T. Ng, S.M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault. 2013. The conll-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria, August. Association for Computational Linguistics.

H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

V. Punyakanok and D. Roth. 2001. The use of classifiers in sequential inference. In *NIPS*.

A. Rozovskaya and D. Roth. 2011. Algorithm selection and model adaptation for esl correction tasks. In *ACL*.

A. Rozovskaya and D. Roth. 2013. Joint learning and inference for grammatical error correction. In *EMNLP*, 10.

A. Rozovskaya, M. Sammons, and D. Roth. 2012. The UI system in the HOO 2012 shared task on error correction. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) Workshop on Innovative Use of NLP for Building Educational Applications*.

A. Rozovskaya, K.-W. Chang, M. Sammons, and D. Roth. 2013. The University of Illinois system in the CoNLL-2013 shared task. In *CoNLL Shared Task*.

A. Rozovskaya, D. Roth, and V. Srikumar. 2014. Correcting grammatical verb errors. In *EACL*.