University of Illinois System in HOO Text Correction Shared Task

Alla Rozovskaya Mark Sammons Joshua Gioja Dan Roth
Cognitive Computation Group
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{rozovska,mssammon,gioja,danr}@illinois.edu

Abstract

In this paper, we describe the University of Illinois system that participated in Helping Our Own (HOO), a shared task in text correction. We target several common errors, such as articles, prepositions, word choice, and punctuation errors, and we describe the approaches taken to address each error type. Our system is based on a combination of classifiers, combined with adaptation techniques for article and preposition detection. We ranked first in all three evaluation metrics (Detection, Recognition and Correction) among six participating teams. We also present type-based scores on preposition and article error correction and demonstrate that our approach achieves best performance in each task.

1 Introduction

The Text Correction task addresses the problem of detecting and correcting mistakes in text. This task is challenging, since many errors are not easy to detect, such as context-sensitive spelling mistakes that involve confusing valid words in a language (e.g. "there" and "their"). Recently, text correction has taken an interesting turn by focusing on context-sensitive errors made by English as a Second Language (ESL) writers. The HOO shared task (Dale and Kilgarriff, 2011) focuses on writing mistakes made by non-native writers of English in the context of Natural Language Processing community.

This paper presents our entry in the HOO shared task. We target several common types of errors using a combination of discriminative and probabilistic classifiers, together with adaptation techniques for article and preposition detection. Our system ranked first in all three evaluation metrics (Detection, Recognition, and Correction). The description of the evaluation schema and the results of the participating teams can be found in Dale and Kilgarriff (2011). We also evaluate the performance of two system components (Sec. 2), those that target article and preposition errors, and compare them to the performance of other teams (Sec. 3).

2 System Components

Our system comprises components that address article and preposition mistakes, word choice errors, and punctuation errors. Table 1 lists the error types that our system targets and shows sample errors from the pilot data[1].

2.1 Article and Preposition Classifiers

We submitted several versions of article and preposition classifiers that build on elements of the systems described in Rozovskaya and Roth (2010b) and Rozovskaya and Roth (2010c).

The systems are trained on the ACL Anthology corpus, which contains 10 million articles and 5 million prepositions[2]; some versions also use additional data from English Wikipedia and the New York Times section of the Gigaword corpus (Linguistic Data Consortium, 2003). Our experiments on the pilot data showed a significant performance gain when training on the ACL Anthology corpus,

[1] The shared task data are split into pilot and test. Each part consists of text fragments from 19 documents, with one fragment from each document included in pilot and one in test.

[2] We consider the top 17 English prepositions.

Component	Relative Freq.	Targeted Errors	Examples
Article	18%	Missing/Unnecessary/ Replacement	Section 5.1 describes the details of ∅*/*the* evaluation metrics. The main advantage of *the*/∅ phonetic alignment is that it requires no training data.
Preposition	9%	Replacement	Pseudo-word searching problem is the same *to*/*as* decomposition of a given sentence into pseudo-words.
Word choice	-	Various lexical and grammatical errors	
Punctuation	18%	Missing/Unnecessary	In the thesaurus we incorporate *LCSbased*/*LCS-based* semantic description for each verb class.

Table 1: **System components.** The column "Relative frequency" shows the the proportion of a given error type in the pilot data. The category "Article" is based on the statistics for determiner errors, the majority of which involve articles.

compared to a system trained on other data, but we observed only a small improvement when other data were added to the ACL Anthology corpus.

The classifiers use features that are based on word n-grams, part-of-speech tags and phrase chunks. The systems use a discriminative learning framework and the regularized version of Averaged Perceptron in Learning Based Java[3] (LBJ, (Rizzolo and Roth, 2007)). This linear learning algorithm is known to be among the best linear learning approaches and has been shown to produce state-of-the-art results on many natural language applications (Punyakanok et al., 2008).

2.1.1 Adaptation to the Error Patterns of the ESL Writers

Mistakes made by non-native speakers are systematic and also depend on the first language of the writer (Lee and Seneff, 2008; Rozovskaya and Roth, 2010a). Injecting knowledge about typical errors into the system improves its performance significantly. While some approaches use this knowledge directly, by training a system on annotated learner data (Han et al., 2010; Gamon, 2010), there is often not enough annotated data for training. In our previous work, we proposed methods to adapt a model to the typical errors of the writers (Rozovskaya and Roth, 2010c; Rozovskaya and Roth, 2010b). The methods use error statistics based only on a small amount of annotation. The preposition and article systems use these methods with additional improvements.

An interesting distinction of the HOO data is that both the pilot and the test fragments are derived from the same set of papers. The size of the pilot data is not sufficient for training a competitive system, but applying the adaptation methods improves the quality of the system by a large margin (Table 2)[4].

System	No adaptation	Adapted
Articles	0.42	0.56
Prepositions	0.38	0.44

Table 2: **Adaptation to the typical errors**. F-score on detection on the pilot data. Error statistics are found in 10-fold cross-validation .

2.2 Word Choice Errors

This component of our system is the most flexible one and does not focus on one type of error but addresses various context-sensitive confusions: spelling errors, grammatical errors, and word choice errors. This component uses a Naïve Bayes classifier trained on the ACL Anthology corpus and the New York Times section of the North American News Text Corpus. The confusion sets include word confusions from the HOO pilot data. The Naïve Bayes formulation allows this component to be flexible with the types of confusions it addresses, unlike the discriminative framework.

2.3 Punctuation Errors

We address two types of punctuation errors, missing commas and misuse of hyphens. We define a set of rules to insert missing commas. Below we describe the hyphen checker.

2.3.1 Hyphen Checker

The hyphen corrector was developed to detect and propose corrections for: 1) inappropriate use of a hyphen to join two words that should be separate tokens; 2) inappropriate use of a hyphen to split

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

[4] The classifiers applied to the test data are adapted using error statistics based on the pilot data.

two words that should be conjoined to form a single word; and 3) omission of a hyphen, resulting in a pair of whitespace-separated words.

We extracted mappings between hyphenated and non-hyphenated sequences using n-gram counts computed from the ACL Anthology corpus by observing the frequency with which the same underlying token sequence occurs either as a single token, as two separate tokens joined by a hyphen, and as two separate tokens with no hyphen. Mappings were extracted for those sequences where one usage was at least 50% more frequent than the others. Discovered rules correct, for example, "paralinguistics" to "paralinguistics" and "pair wise" to "pairwise".

3 Evaluation

The task evaluation uses three metrics, Detection, Recognition, and Correction. In each metric, Recall, Precision and F-score are computed relative to the total number of edits in the corpus (see Dale and Kilgarriff (2011) for a description of the scoring metrics and for the overall ranking of the individual systems). We thought that it would also be interesting to see how the systems compare for two very common error types: articles and prepositions[5]. We have done a comprehensive and slightly different evaluation, computed relative to the edits that involve articles or prepositions, respectively, for each error type[6].

We also evaluate these two tasks by comparing the accuracy of the data before running the system (the "baseline") to the accuracy of the data after running the system. This evaluation shows whether the system reduces or increases the number of errors in the data. The accuracy and the baseline are computed as described in Rozovskaya and Roth (2010c) and the results are shown in Table 5.

[5]Dale and Kilgarriff (2011) show evaluation by error type only for Recall because it is not possible to compute Precision for many other error types. Since it is easy to obtain high recall by proposing many edits (neglecting the precision performance) and, similarly, easy to obtain high precision by just proposing no edits, we present results sorted by F-score rather than by recall and/or precision, as in Dale and Kilgarriff (2011). For the same reason, we also choose the best run of each system based on this measure rather than choosing runs that are doing well just on one of the relevant measures (and, likely very poorly on the other).

[6]For articles, we consider all article edits (see Table 1). For prepositions, replacements involving the top 36 most frequent English prepositions are considered; they account for all preposition replacements made by the participating systems.

Team	Run	Detection	Recognition	Correction
JU	0	0.029	0.029	0.029
LI	3	0.048	0.048	0.033
NU	0	0.372	0.368	0.276
UD	-	-	-	
UI	8	**0.505**	**0.505**	**0.449**
UT	1	0.040	0.025	0.025

Table 3: **Type-based performance: Articles**. For each team, the F-scores for the best run are shown. Results only shown for the teams that address these errors.

Team	Run	Detection	Recognition	Correction
JU	0	0.035	0.035	0.035
LI	8	0.039	0.039	0.039
NU	0	0.266	0.266	0.168
UD	5	0.079	0.079	0.000
UI	8	**0.488**	**0.488**	**0.363**
UT	4	0.202	0.202	0.117

Table 4: **Type-based performance: Prepositions.** For each team, the F-scores for the best run are shown.

Team	Run	Articles	Team	Run	Prepositions
JU	0	0.9280	JU	0	0.9488
LI	3	0.9372	LI	8	0.9546
NU	0	0.9149	NU	0	0.9436
UD	-	-	UD	8	0.9552
UI	5	**0.9424**	UI	9	**0.9562**
UT	7	0.9362	UT	6	0.9372
Baseline		0.9364	Baseline		0.9552

Table 5: **Accuracy results**. "Baseline" is the proportion of correct examples in the data.

4 Conclusion

The shared task is the first competition in text correction, and our team has learned a lot from participating in it – not least, the breadth of error types. We have described the system we entered in the shared task, outlining the approaches we took to address each error type. We also demonstrated the success of our technique for adapting classifiers to writer's errors.

Acknowledgments

The authors thank Jeff Pasternack for his help. This research is supported by a grant from the U.S. Department of Education and is partly supported by the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-018.

References

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*.

M. Gamon. 2010. Using mostly native data to correct errors in learners' writing. In *NAACL*, pages 163–171, Los Angeles, California, June.

N. Han, J. Tetreault, S. Lee, and J. Ha. 2010. Using an error-annotated learner corpus to develop and ESL/EFL error correction system. In *LREC*, Malta, May.

J. Lee and S. Seneff. 2008. An analysis of grammatical errors in non-native speech in English. In *Proceedings of the 2008 Spoken Language Technology Workshop*.

V. Punyakanok, D. Roth, and W. Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2).

N. Rizzolo and D. Roth. 2007. Modeling Discriminative Global Inference. In *Proceedings of the First International Conference on Semantic Computing (ICSC)*, pages 597–604, Irvine, California, September. IEEE.

A. Rozovskaya and D. Roth. 2010a. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*.

A. Rozovskaya and D. Roth. 2010b. Generating confusion sets for context-sensitive error correction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A. Rozovskaya and D. Roth. 2010c. Training paradigms for correcting errors in grammar and usage. In *Proceedings of the NAACL-HLT*.

A. Rozovskaya and D. Roth. 2011. Algorithm selection and model adaptation for esl correction tasks. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, Portland, Oregon, 6. Association for Computational Linguistics.