Robust Systems for Preposition Error Correction Using Wikipedia Revisions

Aoife Cahill*, **Nitin Madnani***, **Joel Tetreault**† **and Diane Napolitano***

* Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA
{acahill, nmadnani, dnapolitano}@ets.org
† Nuance Communications, Inc., 1198 E. Arques Ave, Sunnyvale, CA 94085, USA
Joel.Tetreault@nuance.com

Abstract

We show that existing methods for training preposition error correction systems, whether using well-edited text or error-annotated corpora, do not generalize across very different test sets. We present a new, large error-annotated corpus and use it to train systems that generalize across three different test sets, each from a different domain and with different error characteristics. This new corpus is automatically extracted from Wikipedia revisions and contains over one million instances of preposition corrections.

1 Introduction

One of the main themes that has defined the field of automatic grammatical error correction has been the availability of error-annotated learner data to train and test a system. Some errors, such as determiner-noun number agreement, are easily corrected using rules and regular expressions (Leacock et al., 2010). On the other hand, errors involving the usage of prepositions and articles are influenced by several factors including the local context, the prior discourse and semantics. These errors are better handled by statistical models which potentially require millions of training examples.

Most statistical approaches to grammatical error correction have used one of the following training paradigms: 1) training solely on examples of correct usage (Han et al., 2006); 2) training on examples of correct usage and artificially generated errors (Rozovskaya and Roth, 2010); and 3) training on examples of correct usage and real learner errors (Dahlmeier and Ng, 2011; Dale et al., 2012). The latter two methods require annotated corpora of errors, and while they have shown great promise, manually annotating grammatical errors in a large enough corpus of learner writing is often a costly and time-consuming endeavor.

In order to efficiently and automatically acquire a very large corpus of annotated learner errors, we investigate the use of error corrections extracted from Wikipedia revision history. While Wikipedia revision history has shown promise for other NLP tasks including paraphrase generation (Max and Wisniewski, 2010; Nelken and Yamangil, 2008) and spelling correction (Zesch, 2012), this resource has not been used for the task of grammatical error correction.

To evaluate the usefulness of Wikipedia revision history for grammatical error correction, we address the task of correcting errors in preposition selection (i.e., where the context licenses the use of a preposition, but the writer selects the wrong one). We first train a model directly on instances of correct and incorrect preposition usage extracted from the Wikipedia revision data. We also generate artificial errors using the confusion distributions derived from this data. We compare both of these approaches to models trained on well-edited text and evaluate each on three test sets with a range of different characteristics. Each training paradigm is applied to multiple data sources for comparison. With these multiple evaluations, we address the following research questions:

1. Across multiple test sets, which data source

Proceedings of NAACL-HLT 2013, pages 507–517,
Atlanta, Georgia, 9–14 June 2013.

is more useful for correcting preposition errors: a large amount of well-edited text, a large amount of potentially noisy error-annotated data (either artificially generated or automatically extracted) or a smaller amount of higher quality error-annotated data?

2. Given error-annotated data, is it better to train on the corrections directly or to use the confusion distributions derived from these corrections for generating artificial errors in well-edited text?

3. What is the impact of having a mismatch in the error distributions of the training and test sets?

2 Related Work

In this section, we only review work in preposition error correction in terms of the three training paradigms and refer the reader to Leacock et al. (2010) for a more comprehensive review of the field.

2.1 Training on Well-Edited Text

Early approaches to error detection and correction did not have access to large amounts of error-annotated data to train statistical models and thus, systems were trained on millions of well-edited examples from news text instead (Gamon et al., 2008; Tetreault and Chodorow, 2008; De Felice and Pulman, 2009). Feature sets usually consisted of n-grams around the preposition, POS sequences, syntactic features and semantic information. Since the model only had knowledge of correct usage, an error was flagged if the system's prediction for a particular preposition context differed from the preposition the writer used.

2.2 Artificial Errors

The issue with training solely on correct usage was that the systems had no knowledge of typical learner errors. Ideally, a system would be trained on examples of correct and incorrect usage, however, for many years, such error-annotated corpora were not available. Instead, several researchers generated artificial errors based on the error distributions derived from the error-annotated learner corpora available at the time. Izumi et al. (2003) was the first to evaluate a model trained on incorrect usage as well as artificial errors for the task of correcting several different error types, including prepositions. However, with limited training data, system performance was quite poor. Rozovskaya and Roth (2010) evaluated different ways of generating artificial errors and found that a system trained on artificial errors could outperform the more traditional training paradigm of using only well-edited texts. Most recently, Imamura et al. (2012) showed that performance could be improved by training a model on artificial errors and addressing domain adaptation for the task of Japanese particle correction.

2.3 Error-Annotated Learner Corpora

Recently, error-annotated learner data has become more readily and publicly available allowing models to be trained on both examples of correct usage as well typical learner errors. Han et al. (2010) showed that a preposition error detection and correction system trained on 100,000 annotated preposition errors from the Chungdahm Corpus of Korean Learner English (in addition to 1 million examples of correct usage) outperformed a model trained only on 5 million examples of correct usage. Gamon (2010) and Dahlmeier and Ng (2011) showed that combining models trained separately on examples of correct and incorrect usage could also improve the performance of a preposition error correction system.

3 Mining Wikipedia Revisions for Grammatical Error Corrections

3.1 Related Work

Many NLP researchers have taken advantage of the wealth of information available in Wikipedia revisions. Dutrey et al. (2011) define a typology of modifications found in the French Wikipedia (WiCoPaCo). They show that the kinds of edits made range from specific lexical changes to more general rewrite edits. Similar types of edits are found in the English Wikipedia. The data extracted from Wikipedia revisions has been used for a wide variety of tasks including spelling correction (Max and Wisniewski, 2010; Zesch, 2012), lexical error detection (Nelken and Yamangil, 2008), sentence compression (Yamangil and Nelken, 2008), paraphrase generation (Max and Wisniewski, 2010; Nelken and Yamangil, 2008), lexical simplification (Yatskar et al., 2010) and entailment (Zanzotto and Pennacchiotti, 2010;

(1) [Wiki clean] In addition, sometimes it is also left to stand overnight (*at* → *in*) the refrigerator.

(2) [Wiki clean] Also none of the witnesses present (*of* → *on*) those dates supports Ranneft's claims.

(3) [Wiki dirty] . . . cirque has a permanent production (*to* → *at*) the Mirage, love.

(4) [Wiki dirty] In the late 19th century Vasilli Andreyev a salon violinist took up the balalaika in his performances for French tourists (*in* → *to*) Petersburg.

Figure 1: Example sentences with preposition errors extracted from Wikipedia revisions. The second preposition is assumed to be the correction.

Cabrio et al., 2012). To our knowledge, no one has previously extracted data for training a grammatical error detection system from Wikipedia revisions.

3.2 Extracting Preposition Correction Data from Wikipedia Revisions

As the source of our Wikipedia revisions, we used an XML snapshot of Wikipedia generated in July 2011 containing 8,735,890 articles and 288,583,063 revisions.[1] We then used the following process to extract preposition errors and their corresponding corrections from this snapshot:

Step 1: Extract the plain text versions of all revisions of all articles using the Java Wikipedia Library (Ferschke et al., 2011).

Step 2: For each Wikipedia article, compare each revision with the revision immediately preceding it using an efficient *diff* algorithm.[2]

Step 3: Compute all 1-word **edit chains** for the article, i.e., sequences of related edits derived from all revisions of the same article. For example, say revision 10 of an article inserts the preposition *of* into a sentence and revision 12 changes that preposition to *on*. Assuming that no other revisions change this sentence, the corresponding edit chain would contain the following 3 elements: $\epsilon \rightarrow of \rightarrow on$. The extracted chains contain the full context on either side of the 1-word edit, up to the automatically detected sentence boundaries.

Step 4: (a) Ignore any *circular* chains, i.e., where the first element in the edit chain is the same as the last element. (b) Collapse all *non-circular* chains, i.e., only retain the first and the last elements in a chain. Both these decisions are motivated by the assumption that the intermediate links in the chain are unreliable for training an error correction system since a Wikipedia contributor modified them.

Step 5 : From all remaining 2-element chains, find those where a preposition is replaced with another preposition. If the preposition edit is the only edit in the sentence, we convert the chain into a sentence pair and label it *clean*. If there are other 1-word edits but not within 5 words of the preposition edit on either side, we label the sentence *somewhat clean*. Otherwise, we label it *dirty*. The motivation is that the presence of other nearby edits make the preposition correction less reliable when used in isolation, due to the possible dependencies between corrections.

All extracted sentences were part-of-speech tagged using the Stanford Tagger (Toutanova et al., 2003). Using the above process, we are able to extract approximately 2 million sentences containing prepositions errors and their corrections. Some examples of the sentences we extracted are given in Figure 1. Example (4) shows an example of a bad correction.

4 Corpora

We use several corpora for training and testing our preposition error correction system. The properties of each are outlined in Table 1, organized by paradigm. For each corpus we report the total number of prepositions used for training, as well as the number and percentage of preposition corrections.

4.1 Well-edited Text

We train our system on two well-edited corpora. The first is the same corpus used by Tetreault and

[1] http://dumps.wikimedia.org/enwiki/

[2] http://code.google.com/p/google-diff-match-patch/

	Corpus	Total # Preps	# Corrected Preps	
Well-edited Text	Wikipedia Snapshot (10m sents)	26,069,860	0	(0%)
	Lexile/SJM	6,719,077	0	(0%)
Artificially Generated Errors	Wikipedia Snapshot	26,127,464	2,844,227	(10.9%)
	Lexile/SJM	6,723,206	792,195	(11.8%)
Naturally Occurring Errors	Wikipedia Revisions All	7,125,317	1,027,643	(20.6%)
	Wikipedia Revisions ∼Clean	3,001,900	381,644	(12.7%)
	Wikipedia Revisions Clean	1,978,802	266,275	(14.4%)
	Lang-8	129,987	53,493	(41.2%)
	NUCLE Train	72,741	922	(1.3%)
Test Corpora	NUCLE Test	9,366	125	(1.3%)
	FCE	33,243	2,900	(8.7%)
	HOO 2011 Test	1,703	81	(4.8%)

Table 1: Corpora characteristics

Chodorow (2008), comprising roughly 1.8 million sentences from the San Jose Mercury News Corpus[3] and roughly 1.8 million sentences from grades 11 and 12 of the MetaMetrics Lexile Corpus. Our second corpus is a random sample of 10 million sentences containing at least one preposition from the June 2012 snapshot of English Wikipedia Articles.[4]

4.2 Artificially Generated Errors

Similar to Foster and Andersen (2009) and Rozovskaya and Roth (2010), we artificially introduce preposition errors into well-edited corpora (the two described above). We do this based on a distribution of possible confusions and train a model that is aware of the corrections. The two sets of confusion distributions we used were derived based on the errors extracted from Wikipedia revisions and Lang-8 respectively (discussed in Section 4.3). For each corrected preposition p_i in the revision data, we calculated $P(p_i|p_j)$, where p_j is each of the possible original prepositions that were confused with p_i. Then, for each sentence in the well-edited text, all prepositions are extracted. A preposition is randomly selected (without replacement) and changed based on the distribution of possible confusions (note that the original preposition is also included in the distribution, usually with a high probability, meaning that there is a strong preference not to change the preposition). If a preposition is changed to something other than the original preposition, all remaining prepositions in the sentence are left unchanged.

4.3 Naturally Occurring Errors

We have a number of corpora that contain annotated preposition errors. Note that we are only considering incorrectly selected prepositions, we do not consider missing or extraneous.

NUCLE The NUS Corpus of Learner English (NUCLE)[5] contains one million words of learner essay text, manually annotated with error tags and corrections. We use the same training, dev and test splits as Dahlmeier and Ng (2011).

FCE The CLC FCE Dataset[6] is a collection of 1,244 exam scripts written by learners of English as part of the Cambridge ESOL First Certificate in English (Yannakoudakis et al., 2011). It includes demographic metadata about the candidate, a grade for each essay and manually-annotated error corrections.

Wikipedia We use three versions of the preposition errors extracted from the Wikipedia revisions as described in Section 3.2. The first includes corrections where the preposition was the only word corrected in the entire sentence

[3]The San Jose Mercury News is available from the Linguistic Data Consortium (catalog number LDC93T3A).

[4]We used a newer version of the Wikipedia text for the well-edited text, since we assume that more recent versions of the text will be most grammatical, and therefore closer to well-edited.

[5]http://bit.ly/nuclecorpus

[6]http://ilexir.co.uk/applications/clc-fce-dataset/

(*clean*). The second contains all *clean* corrections, as well as all corrections where there were no other edits within a five-word span on either side of the preposition (∼*clean*). The third contains all corrections regardless of any other changes in the surrounding context (*all*).

Lang-8 The Lang-8 website contains journals written by language learners, where native speakers highlight and correct errors on a sentence-by-sentence basis. As a result, it contains typical grammatical mistakes made by language learners, which can be easily downloaded. We automatically extract 75,622 sentences with preposition errors and corrections from the first million journal entries.[7]

HOO 2011 We take the test set from the HOO 2011 shared task (Dale and Kilgarriff, 2011) and extract all examples of preposition selection errors. The texts are fragments of ACL papers that have been manually annotated for grammatical errors.[8]

It is important to note that the three test sets we use are from entirely different domains: exam scripts from non-native English speakers (FCE), essays by highly proficient college students in Singapore (NUCLE) and ACL papers (HOO). In addition, they have a different number of total prepositions as well as erroneous prepositions.

5 Preposition Error Correction Experiments

We use the preposition error correction model described in Tetreault and Chodorow (2008)[9] to evaluate the many ways of using Wikipedia error corrections as described in the Section 4. We use this system since it has been recreated for other work (Dahlmeier and Ng, 2011; Tetreault et al., 2010) and is similar in methodology to Gamon et al. (2008) and De Felice and Pulman (2009). In short, the method models the problem of preposition error correction (for replacement errors) as a 36-way classification problem using a multinomial logistic regression model.[10] The system uses 25 lexical, syntactic and n-gram features derived from the contexts of each preposition training instance.

We modified the training paradigm of Tetreault and Chodorow (2008) so that a model could be trained on examples of correct usage as well as actual errors. We did this by adding a new feature specifying the writer's original preposition (as in Han et al. (2010) and Dahlmeier and Ng (2011)).

5.1 Results

We train a preposition correction system using each of the three data paradigms and test on the FCE, NUCLE and HOO 2011 test corpora. For each preposition in the test corpus, we record whether the system predicted that it should be changed, and if so, what it should be changed to. We then compare the prediction to the annotation in the test corpus. We report results in terms of f-score, where precision and recall are calculated as follows:[11]

$$\text{Precision} = \frac{\text{Number of correct preposition corrections}}{\text{Total number of corrections suggested}}$$

$$\text{Recall} = \frac{\text{Number of correct preposition corrections}}{\text{Total number of corrections in test set}}$$

Note that due to the high volume of unchanged prepositions in the test corpus, we obtain very high accuracies, which are not indicative of true performance, and are not included in our results.

The results of our experiments are presented in Table 2.[12] The first part of the table shows the f-scores of preposition error correction systems that

[7] Tajiri et al. (2012) extract a corpus of English verb phrases corrected for tense/aspect errors from Lang-8. They kindly provided us with their scripts to carry out the scraping of Lang-8.

[8] The results of the HOO 2011 shared task were not reported at level of preposition selection error, therefore it is not possible to compare the results presented in this paper with those results.

[9] Note that in that work, the model was evaluated in terms of preposition error *detection* rather than correction, however the model itself does not change.

[10] We use liblinear (Fan et al., 2008) with the L1-regularized logistic regression solver and default parameters.

[11] As Chodorow et al. (2012) note, it is not clear how to handle cases where the system predicts a preposition that is neither the same as the writer preposition nor the correct preposition. We count these cases as false positives.

[12] No thresholds were used in the systems that were trained on well-edited text. Traditionally, thresholds are applied so as to only predict a correction when the system is highly confident. This has the effect of increasing precision at the cost of recall, and sometimes leads to an overall improved f-score. Here we take the prediction of the system, regardless of the confidence, reflecting a lower-bound of this method.

	Data Source	Paradigm	CLC-FCE N=33,243	NUCLE N=9,366	HOO2011 N=1,703
Without Wikipedia Revisions (**nonWikiRev**)	Wikipedia Snapshot	Well-edited Text	24.43*	5.02*	12.36*
	Lexile/SJM	Well-edited Text	24.73*	4.29*	9.73*
	Wikipedia Snapshot	Artificial Errors (Lang-8)	42.15*	19.91*	28.75
	Lexile/SJM	Artificial Errors (Lang-8)	**45.36**	18.00*	25.15
	Lang-8	Error-annotated Text	38.22*	8.18*	24.00
	NUCLE train	Error-annotated Text	5.38*	20.14	4.82*
With Wikipedia Revisions (**WikiRev**)	Wikipedia Snapshot	Artificial Errors (Wiki)	31.17*	24.52	28.30
	Lexile/SJM	Artificial Errors (Wiki)	34.35*	23.38	32.76
	Wikipedia Revisions All	Error-annotated Text	33.59*	**26.39**	**36.84**
	Wikipedia Revisions ∼Clean	Error-annotated Text	29.68*	22.13	36.04
	Wikipedia Revisions Clean	Error-annotated Text	28.09*	21.74	28.30

Table 2: Preposition selection error correction results (f-score). The systems with scores in bold are statistically significantly better than all systems marked with an asterisk ($p < 0.01$). Confidence intervals were obtained using bootstrap resampling with 50,000 replicates.

one might be able to train with publicly available data excluding the Wikipedia revisions that we have extracted. We refer to these systems as **nonWikiRev** systems. The second part of the table shows the f-scores of systems trained on the Wikipedia revisions data – either directly on the annotated errors or on the artificial errors produced using the confusion distributions derived from these annotated errors. We refer to this second set of systems as **WikiRev** systems. The **nonWikiRev** systems perform inconsistently, heavily dependent on the characteristics of the test set in question. On the other hand, it is obvious that the **WikiRev** systems — while not always outperforming the best **nonWikiRev** systems — generalize much better across the three test sets. In fact, for the NUCLE test set, the best **WikiRev** system performs as well as the **nonWikiRev** system trained on data from the *same* domain and with *identical* error characteristics as the test set. The distributions of errors in the three test sets are not similar, and therefore, the stability in performance of the **WikiRev** systems cannot be attributed to the hypothesis that the **WikiRev** training data error distributions are more similar to the test data than any of the other training corpora. Therefore, we claim that if a preposition error correction system is to be deployed on data for which the error characteristics are not known in advance, i.e. most real-world scenarios, training the system using Wikipedia revisions is likely to be the most robust option.

6 Discussion

We examine the results of our experiments in light of the research questions we posed in Section 1.

6.1 Which Data Source is More Useful?

We wanted to know whether it was better to have a smaller corpus of carefully annotated corrections, or a much larger (but automatically generated, and therefore noisier) error-annotated corpus. We also wanted to compare this scenario to training on large amounts of well-edited text. From our experiments, it is clear that the composition of the test set plays a major role in answering this question. On a test set with few corrections (NUCLE), training on well-edited text (and without using thresholds) performs particularly poorly. On the other hand, when evaluating on the FCE test set which contains far more errors, training on well-edited text performs reasonably well (though statistically significantly worse than training on all of the Wikipedia errors). Similarly, training on the smaller, high-quality NUCLE corpus and evaluating on the NUCLE test set achieves good results, however training on NUCLE and testing on FCE achieves the lowest f-score of all our systems on that test set.

Figure 2 shows the learning curves obtained by increasing the size of the training data for two of the test sets.[13] Although one might assume

[13]For space reasons, the graph for HOO2011 is omitted. Also note that the results in Table 2 may not appear in the graph,

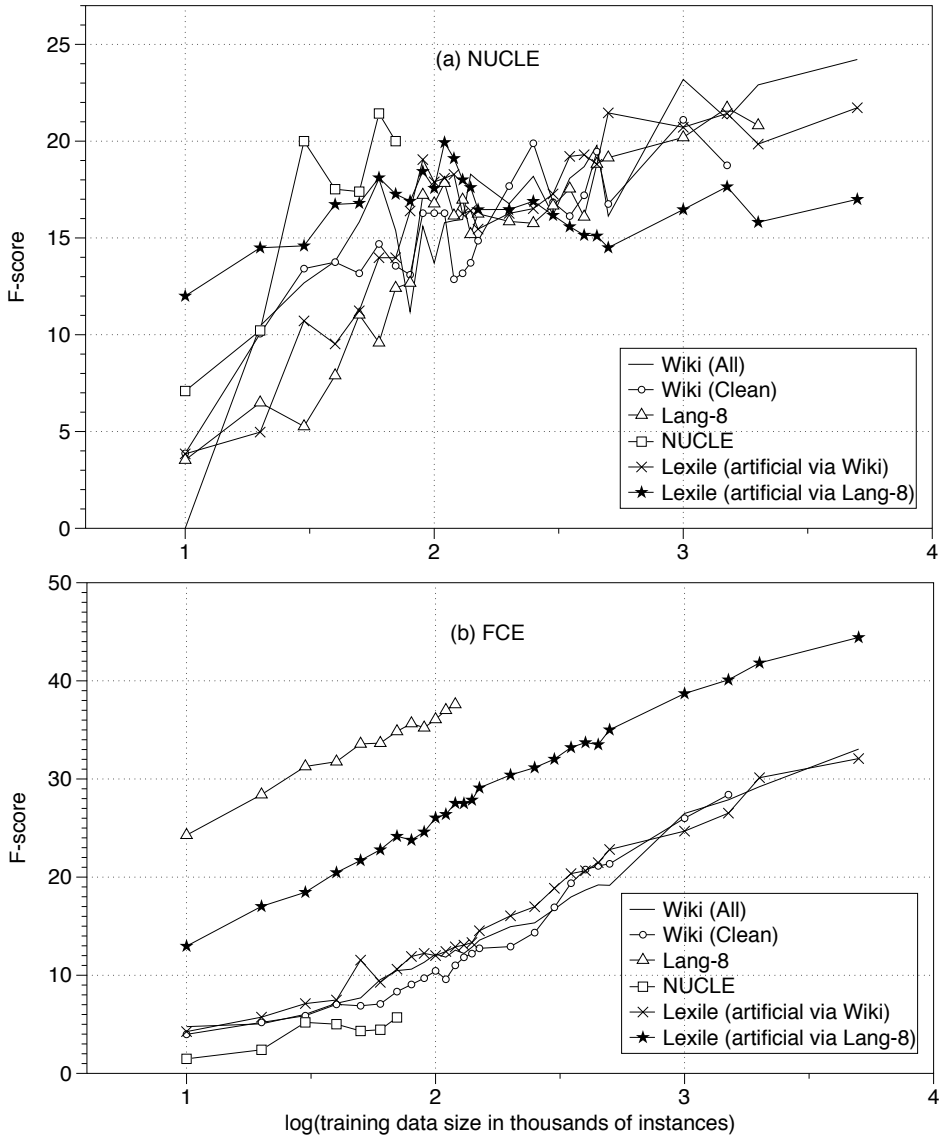

Figure 2: The effect of varying the size of the training corpus

that Wikipedia-*clean* would be more reliable than Wikipedia-*all*, the *cleanness* of the Wikipedia data seems to make very little difference, probably because the data extracted in the *dirty* contexts is not as noisy as we expected. Interestingly, it also seems that additional data would lead to further improvements for models trained on artificial errors in Lexile data and for those trained on all of the automatically extracted Wikipedia errors.

Another interesting aspect of Figure 2 is that training on the Lang-8 data shows a very steep rising trend. This suggests that automatically-scraped data that is highly targeted towards language learners is very useful in correcting preposition errors in texts where they are reasonably frequent.

6.2 Natural or Artificially Generated Errors?

Table 2 shows that training on artificially generated errors via Wikipedia revisions performs fairly consistently across test corpora. While using Lang-8 for artificial error generation is also quite promising for FCE, it does not generalize across test sets.

since we were sampling at specific data points which did not correspond exactly to the total sizes of the training corpora.

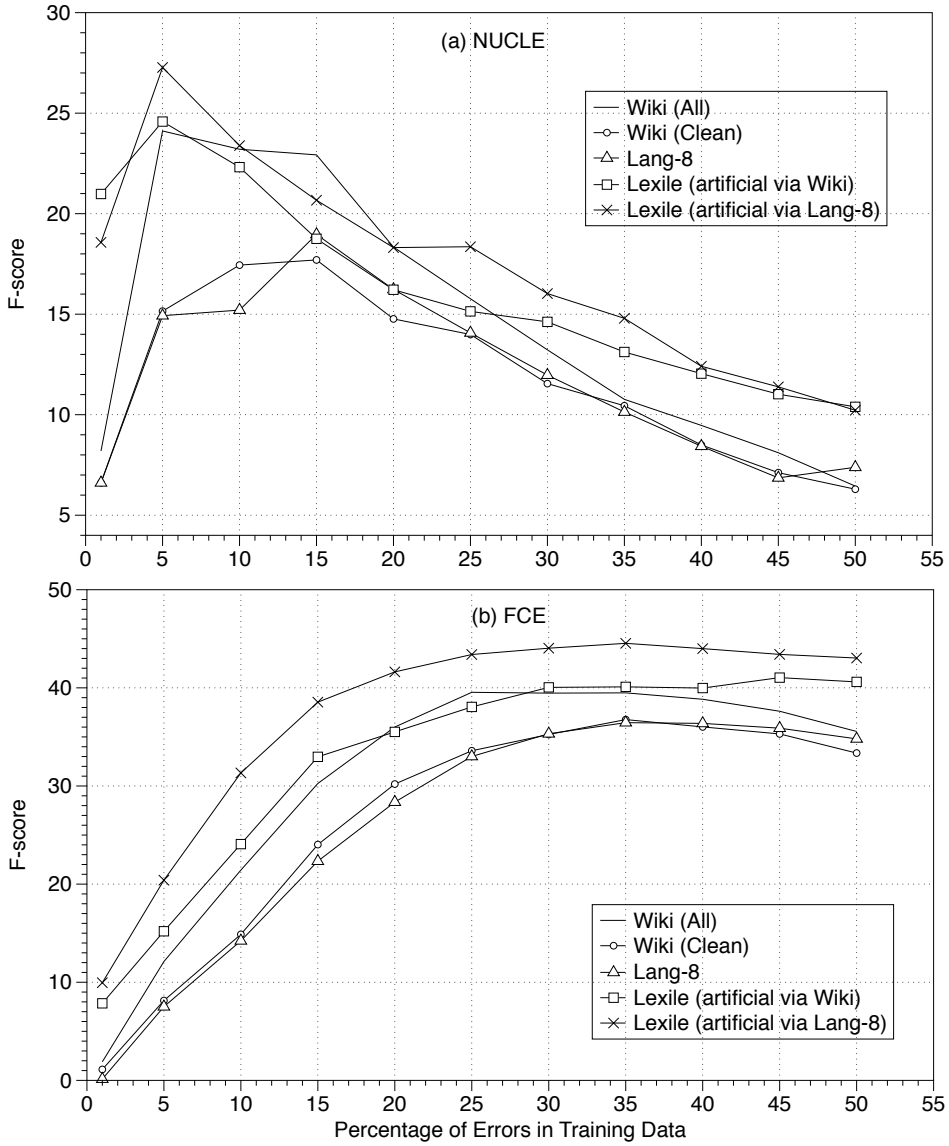

Figure 3: The effect of varying the percentage of errors in the training corpus

On FCE it achieves the highest results, on NUCLE it performs statistically significantly worse than the best system, and on HOO 2011 it achieves a lower (though not statistically significant) result than the best system. This highlights that extracting errors from Wikipedia is useful in two ways: (1) training a system on the errors alone works well and (2) generating artificial errors in well-edited corpora of different domains and training a system on that also works well. It also indicates that if the system were to be applied to a specific domain, applying the confusion distributions to a domain specific corpus – if available – would likely yield the best results.

6.3 Mismatching Distributions

The proportion of errors in the training and test data plays an important role in the performance of any preposition error correction system. This is clearly evident by comparing system performances across the three test sets which have fairly different compositions. FCE contains a much higher proportion of errors than NUCLE, and HOO falls somewhere in between. Interestingly, the system trained on Lang-8 data (which contains the highest proportion of er-

rors among all training corpora) performs best on the FCE data. On the other hand, the same system performs poorly on NUCLE test which contains far fewer errors. In this instance, the system learns to predict an incorrect preposition too often. We see a similar pattern with the system trained on the NUCLE training data. It performs poorly on FCE which contains many errors, but well on NUCLE test which contains a similar proportion of errors.

In order to better understand the relationship between the percentage of errors in the training data and system performance, we vary the percentage of errors in each training corpus from 1-50% and test on the unchanged FCE and NUCLE test corpora. For each training corpus, we reduce the size to be twice the size of the total number of errors.[14] Keeping this size constant, we then artificially change the percentage of errors. Note that because the total size of the corpus has changed, the results in Table 2 may not appear in the graph. Figure 3 shows the effect on f-score when the data composition is changed. For both test sets, there is a peak after which increasing the proportion of errors in the training corpus is detrimental. For NUCLE test with its low number of preposition errors, this peak is very pronounced. For FCE, it is more of a gentle degradation in performance, but the pattern is clear. Also noteworthy is the fact that the degradation for models trained on artificial errors is less steep suggesting that they may be more stable across test sets.

In general, these results indicate that when building a preposition error detection using error-annotated data, the characteristics of the data to which the system will be applied should play a vital role in how the system is to be trained. Our results show that the **WikiRev** systems are robust across test sets, however if the exact distribution of errors in the data is known in advance, other models may perform better.

7 Conclusion

Although previous approaches to preposition error correction using either well-edited text or small hand-annotated corrections performed well on some specific test set, they did not generalize well across very different test sets. In this paper, we present work that automatically extracts preposition error corrections from Wikipedia Revisions and uses it to build *robust* error correction systems. We show that this data is useful for two purposes. Firstly, a model trained directly on the corrections performs well across test sets. Secondly, models trained on artificial errors generated from the distribution of confusions in the Wikipedia data perform equally well. The distribution of confusions can also be applied to other well-edited corpora in different domains, providing a very powerful method of automatically generating error corpora. The results of our experiments also highlight the importance of the distribution of expected errors in the test set. Models that perform well on one kind of distribution may not necessarily work on a completely different one, as evident in the performances of the systems trained on either Lang-8 or NUCLE. In general, the **WikiRev** models perform well across distributions. We also conducted some preliminary system combination experiments and found that while they yielded promising results, further investigation is necessary. We have also made the Wikipedia preposition correction corpus available for download.[15]

In future work, we will examine whether the results we obtain for English generalize to other Wikipedia languages. We also plan to extract multi-word corrections for other types of errors and to examine the usefulness of including error contexts in our confusion distributions (e.g., preposition confusions following verbs versus those following nouns).

Acknowledgments

The authors would like to thank Daniel Dahlmeier, Torsten Zesch, Mamoru Komachi, Tajiri Toshikazu, Tomoya Mizumoto and Yuji Matsumoto for providing scripts and data that enabled us to carry out this research. We would also like to thank Martin Chodorow and the anonymous reviewers for their helpful suggestions and comments.

References

Elena Cabrio, Bernardo Magnini, and Angelina Ivanova. 2012. Extracting Context-Rich Entailment Rules from

[14] We omit the NUCLE train corpus from this comparison, because it contains too few errors to obtain a meaningful result.

[15] http://bit.ly/etsprepdata

Wikipedia Revision History. In *Proceedings of the 3rd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP*, pages 34–43, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel Tetreault. 2012. Problems in Evaluating Grammatical Error Detection Systems. In *Proceedings of COLING 2012*, pages 611–628, Mumbai, India, December. The COLING 2012 Organizing Committee.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical Error Correction with Alternating Structure Optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 915–923, Portland, Oregon, USA, June. Association for Computational Linguistics.

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France, September. Association for Computational Linguistics.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada, June. Association for Computational Linguistics.

Rachele De Felice and Stephen G. Pulman. 2009. Automatic detection of preposition errors in learner writing. *CALICO Journal*, 26(3):512–528.

Camille Dutrey, Houda Bouamor, Delphine Bernhard, and Aurélien Max. 2011. Local modifications and paraphrases in Wikipedias revision history. *SEPLN journal(Revista de Procesamiento del Lenguaje Natural)*, 46:51–58.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. 2011. Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*.

Jennifer Foster and Oistein Andersen. 2009. GenERRate: Generating Errors for Use in Grammatical Error Detection. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–90, Boulder, Colorado, June. Association for Computational Linguistics.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alex Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 449–456, Hyderabad, India.

Michael Gamon. 2010. Using Mostly Native Data to Correct Errors in Learners' Writing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 163–171, Los Angeles, California, June. Association for Computational Linguistics.

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

Na-Rae Han, Joel Tetreault, Soo-Hwa Lee, and Jin-Young Ha. 2010. Using Error-Annotated ESL Data to Develop an ESL Error Correction System. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Malta.

Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. Grammar Error Correction Using Pseudo-Error Sentences and Domain Adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 388–392, Jeju Island, Korea, July. Association for Computational Linguistics.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003. Automatic Error Detection in the Japanese Learners' English Spoken Data. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 145–148, Sapporo, Japan, July. Association for Computational Linguistics.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan Claypool.

Aurélien Max and Guillaume Wisniewski. 2010. Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Rami Nelken and Elif Yamangil. 2008. Mining Wikipedias Article Revision History for Training

Computational Linguistics Algorithms. In *Proceedings of the 1st AAAI Workshop on Wikipedia and Artificial Intelligence*, pages 31–36, Chicago, IL.

Alla Rozovskaya and Dan Roth. 2010. Generating Confusion Sets for Context-Sensitive Error Correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 961–970, Cambridge, MA, October. Association for Computational Linguistics.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and Aspect Error Correction for ESL Learners Using Global Context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers*, pages 198–202, Jeju Island, Korea.

Joel R. Tetreault and Martin Chodorow. 2008. The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 865–872, Manchester, UK.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. Using Parse Features for Preposition Selection and Error Detection. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 353–358, Uppsala, Sweden, July. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of NAACL*, pages 173–180.

Elif Yamangil and Rani Nelken. 2008. Mining Wikipedia Revision Histories for Improving Sentence Compression. In *Proceedings of ACL-08: HLT, Short Papers*, pages 137–140, Columbus, Ohio, June. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, June. Association for Computational Linguistics.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368, Los Angeles, California, June. Association for Computational Linguistics.

Fabio Massimo Zanzotto and Marco Pennacchiotti. 2010. Expanding textual entailment corpora from Wikipedia using co-training. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 28–36, Beijing, China, August. Coling 2010 Organizing Committee.

Torsten Zesch. 2012. Measuring Contextual Fitness Using Error Contexts Extracted from the Wikipedia Revision History. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 529–538, Avignon, France, April. Association for Computational Linguistics.