An Explicit Feedback System for Preposition Errors based on Wikipedia Revisions

Nitin Madnani and **Aoife Cahill**
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
{nmadnani,acahill}@ets.org

Abstract

This paper presents a proof-of-concept tool for providing automated explicit feedback to language learners based on data mined from Wikipedia revisions. The tool takes a sentence with a grammatical error as input and displays a ranked list of corrections for that error along with evidence to support each correction choice. We use lexical and part-of-speech contexts, as well as query expansion with a thesaurus to automatically match the error with evidence from the Wikipedia revisions. We demonstrate that the tool works well for the task of preposition selection errors, evaluating against a publicly available corpus.

1 Introduction

A core feature of learning to write is receiving feedback and making revisions based on that feedback (Biber et al., 2011; Lipnevich and Smith, 2008; Truscott, 2007; Rock, 2007). In the field of second language acquisition, the main focus has been on *explicit* or *direct* feedback vs. *implicit* or *indirect* feedback. In writing, explicit or direct feedback involves a clear indication of the location of an error as well as the correction itself, or, more recently, a meta-linguistic explanation (of the underlying grammatical rule). Implicit or indirect written feedback indicates that an error has been made at a location, but it does not provide a correction.

The work in this paper describes a novel tool for presenting language learners with explicit feedback based on human-authored revisions in Wikipedia. Here we describe the proof-of-concept tool that provides explicit feedback on one specific category of grammatical errors, preposition selection. We restrict the scope of the tool in order to be able to carry out a focused study, but expect that our findings presented here will also generalize to other error types. The task of preposition selection errors has been well studied (Tetreault and Chodorow, 2008; De Felice and Pulman, 2009; Tetreault et al., 2010; Rozovskaya and Roth, 2010; Dahlmeier and Ng, 2011; Seo et al., 2012; Cahill et al., 2013), and the availability of public, annotated corpora containing such errors provides easy access to evaluation data.

Our tool takes a sentence with a grammatical error as input, and returns a *ranked* list of possible corrections. The tool makes use of frequency of correction in edits to Wikipedia articles (as recorded in the Wikipedia revision history) to calculate the rank order. In addition to the ranked list of suggestions, the tool also provides evidence for each correction based on the actual changes made between different versions of Wikipedia articles. The tool uses the notion of "context similarity" to determine whether a particular edit to a Wikipedia article can provide evidence of a correction in a given context.

Specifically, this paper makes the following contributions:

1. We build a tool to provide explicit feedback for preposition selection errors in the form of ranked lists of suggested corrections.
2. We use evidence from human-authored corrections for each suggested correction on a list.
3. We conduct a detailed examination of how the performance of the tool is affected by varying the type and size of contextual information and by the use of query expansion.

The remainder of this paper is organized as follows: §2 describes related work and §3 outlines potential approaches for using Wikipedia revision data in a feedback tool. §4 outlines the core system

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 79–88,
Baltimore, Maryland USA, June 26, 2014.

for generating feedback and §5 presents an empirical evaluation of this system. In §6 we describe a method for enhancing the system using query expansions. We discuss our findings and some future work in §7 and, finally, conclude in §8.

2 Related Work

Attali (2004) examines the general effect of feedback in the Criterion system (Burstein et al., 2003) and finds that students presented with feedback are able to improve the overall quality of their writing, as measured by an automated scoring system. This study does not investigate different kinds of feedback, but rather looks at the issue of whether feedback in general is useful for students. Shermis et al. (2004) look at groups of students who used Criterion and students who did not and compare their writing performance as measured by high-stakes state assessment. They found that, in general, the students who made use of Criterion and its feedback improved their writing skills. They analyze the distributions of the individual grammar and style error types and found that Criterion helped reduce the number of repeated errors, particularly for mechanics (e.g. spelling and punctuation errors). Chodorow et al. (2010) describe a small study in which Criterion provided feedback about article errors to students writing an essay for a college-level course. They find, similarly to Attali (2004), that the number of article errors was reduced in the final revised version of the essay.

Gamon et al. (2009) describe *ESL Assistant* — a web-based proofreading tool designed for language learners who are native speakers of East-Asian languages. They used a decision-tree approach to detect and offer suggestions for potential article and preposition errors. They also allowed the user to compare the various suggestions by showing results of corresponding web searches. Chodorow et al. (2010) also describe a small study where *ESL Assistant* was used to offer suggestions for potential grammatical errors to web users while they were composing email messages. They reported that users were able to make effective use of the explicit feedback for that task. The tool had been offered as a web service but has since been discontinued.

Our tool is similar to *ESL Assistant* in that both produce a list of possible corrections. The main difference between the tools is that ours automatically derives the ranked list of correction suggestions from a very large corpus of annotated errors, rather than performing a web search on all possible alternatives in the context. The advantage of using an error-annotated corpus is that it contains implicit information about frequent confusion pairs (e.g. "at" instead of "in") that are independent of the frequency of the preposition and the current context.

Milton and Cheng (2010) describe a toolkit for helping Chinese learners of English become more independent writers. The toolkit gives the learners access to online resources including web searches, online concordance tools, and dictionaries. Users are provided with snapshots of the word or structure in context. In Milton (2006), 500 revisions to 323 journal entries were made using an earlier version of this tool. Around 70 of these revisions had misinterpreted the evidence presented or were careless mistakes; the remaining revisions resulted in more natural sounding sentences.

3 Wikipedia Revisions

Our goal is to build a tool that can provide explicit feedback about errors to writers. We take advantage of the recently released Wikipedia preposition error corpus (Cahill et al., 2013) and design our tool based on this large corpus containing sentences annotated for preposition errors and their corrections. The corpus was produced automatically by mining a total of 288 million revisions for 8.8 million articles present in a Wikipedia XML snapshot from 2011. The Wikipedia error corpus, as we refer to in the rest of the paper, contains 2 million sentences annotated with preposition errors and their respective corrections.

There are two possible approaches to building an explicit feedback tool for preposition errors based on this corpus:

1. **Classifier-based**. We could train a classifier on the Wikipedia error corpus to predict the correct preposition in a given context, as Cahill et al. (2013) did. Although this would allow us to suggest corrections for contexts that are unseen in the Wikipedia data, the suggestions would likely be quite noisy given the inherent difficulty of a classification problem with a large number of classes.[1] In addition, this approach would not facilitate pro-

[1] Cahill et al. (2013) used a list of 36 prepositions as classes.

viding evidence for each correction to the user.

2. **Corpus-based**. We could use the Wikipedia error corpus *directly* for feedback. Although this means that suggestions can only be generated for contexts occurring in the Wikipedia data, it also means that all suggestion would be grounded in actual revisions made by other humans on Wikipedia.

We believe that anchoring suggestions to human-authored corrections affords greater utility to a language learner, in line with the current practice in lexicography that emphasizes authentic usage examples (Collins COBUILD learner's dictionary, Sketch Engine (Kilgarriff et al., 2004)). Therefore, in this paper, we choose the second approach to build our tool.

4 Methodology

In order to use the Wikipedia error corpus directly for feedback, we first index the sentences in the corpus using the following fields:

- The incorrect preposition.
- The correct preposition.
- The words, bigrams, and trigrams before (and after) the preposition error (indexed separately).
- The part-of-speech tags, tag bigrams, and tag trigrams before (and after) the error (indexed separately).
- The title and URL of the Wikipedia article in which the sentence occurred.
- The ID of the article revision containing the preposition error.
- The ID of the article revision in which the correction was made.

Once the index is constructed, eliciting explicit feedback is straightforward. The input to the system is a tokenized sentence with a marked up preposition error (e.g. from an automated preposition error detection system). For each input sentence, the Wikipedia index is then searched with the identified preposition error and the words (or n-grams) present in its context. The index returns a list of the possible corrections occurring in the given context. The tool then counts how often each possible preposition is returned as a possible correction and orders its suggestions from most frequent to least frequent. In addition, the tool also displays five randomly chosen sentences from the index as evidence for each correction in order to help the learner make a better choice. The tool can use either the lexical n-grams (n=1,2,3) or the part-of-speech n-grams (n=1,2,3) around the error for the contextualized search of the Wikipedia index.

Figure 1 shows a screenshot of the tool in operation. The input sentence is entered into the text box at the top, with the preposition error enclosed in asterisks. In this case, the tool is using parts-of-speech on either side of the error for context. By default, the tool shows the top five possible corrections as a bar chart, sorted according to how many times the erroneous preposition was changed to the correction in the Wikipedia revision index. In this example, the preposition *of* with the left context of `<DT, NNS>` and the right context of `<DT, NN>` was changed to the preposition *in* 242 times in the Wikipedia revisions. When the user clicks on a bar, the box on the top shows the sentence with the change and the gray box on the right shows 5 (randomly chosen) actual sentences from Wikipedia where the change represented by the bar was made.

If parts-of-speech are chosen as context, the tool uses WebSockets to send the sentence to the Stanford Tagger (Toutanova et al., 2003) in the background and compute its part-of-speech tags before searching the index.

5 Evaluation

In order to determine how well the tool performs at suggesting corrections, we used sentences containing preposition errors from the **CLC FCE dataset**. The CLC FCE Dataset 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). Our evaluation set consists of 3,134 sentences, each containing a single preposition error.

We evaluate the tool on two criteria:

- **Coverage**. We define coverage as the proportion of errors for which the tool is able to suggest any corrections.
- **Accuracy**. The obvious definition of accu-

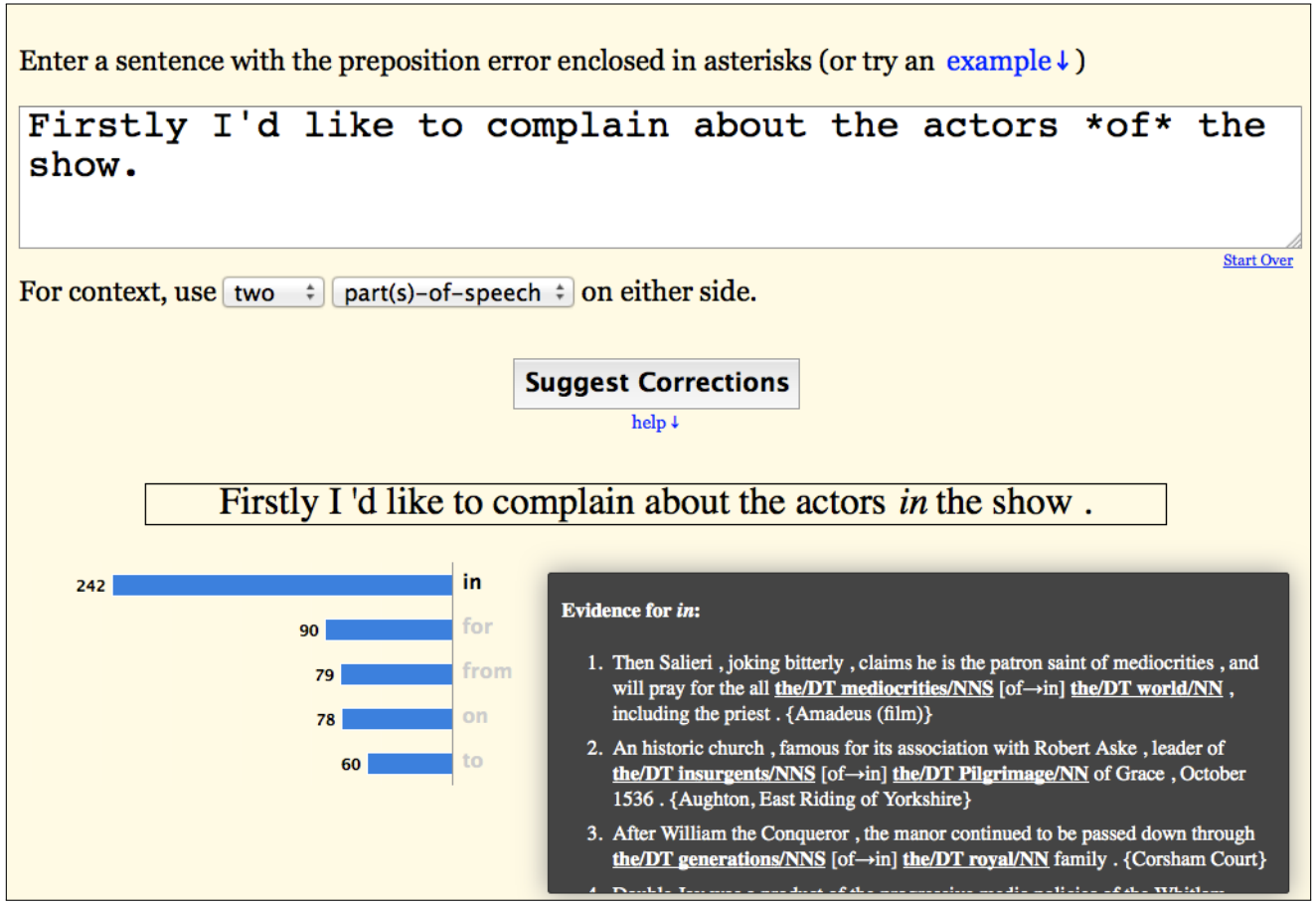

Figure 1: A screenshot of the tool suggesting the top 5 corrections for a sentence using two parts-of-speech on either side of the marked error as context. The corrections are displayed in ranked fashion as a histogram and clicking on one displays the "corrected" sentence above and the corresponding evidence from Wikipedia revisions on the left.

Context	Found	Missed	Blank	MRR
words1	889 (28.4%)	356 (11.4%)	1889 (60.3%)	.522
words2	55 (1.8%)	22 (0.7%)	3057 (97.5%)	.619
words3	16 (0.5%)	5 (0.2%)	3113 (99.3%)	.762
tags1	2821 (90.0%)	241 (7.7%)	72 (2.3%)	.419
tags2	1896 (60.5%)	718 (22.9%)	520 (16.6%)	.390
tags3	661 (21.1%)	633 (20.2%)	1840 (58.7%)	.325

Table 1: A detailed breakdown of the **Found**, **Missing** and **Blank** classes along with the Mean Reciprocal Rank (MRR) values, for different types (words, tags) and sizes (1, 2, or 3 around the error) of contextual information used in the search.

racy would be the proportion of errors for which the tool's best suggestion is the correct one. However, since the tool returns a ranked list of suggestions, it is important to award partial credit for errors where the tool made a correct suggestion but it was not ranked at the top. Therefore, we use the Mean Reciprocal Rank (MRR), a standard metric used for evaluating ranked retrieval systems (Voorhees, 1999). MRR is computed as follows:

$$\text{MRR} = \frac{1}{|S|}\sum_{i=1}^{|S|}\frac{1}{R_i}$$

where S denotes the set of sentences for which ranked lists of suggestions are generated and R_i denotes the rank of the true correction in the list of suggestions the tool returns for sentence i. A higher MRR is better since that means that the tool ranked the true correction closer to the top of the list.

To conduct the evaluation on the FCE dataset, we run each of the sentences through the tool and extract the top 5 suggestions for each error annotated in the sentence.[2] At this point, each error instance input to the tool can be classified as one of three classes:

1. **Found**. The true correction for the error was found in the ranked list of suggestions made by the tool.

2. **Missing**. The true correction for the error was *not* found in the ranked list of suggestions.

3. **Blank**. The tool did not return any suggestions for the error.

[2]In this paper, we separate the tasks of error detection and correction and use the gold standard as an oracle to detect errors and then use our system to propose and rank corrections.

First, we examine the distribution of the three classes across the types and sizes of the contextual information used to conduct the search. Table 1 shows, for each context type and size, a detailed breakdown of the distribution of the three classes along with the mean reciprocal rank (MRR) values.[3] We observe that, with words as contexts, using larger contexts certainly produces more accurate results (as indicated by the larger MRR values). However, we also observe that employing larger contexts reduces coverage (as indicated by the decreasing percentage of **Found** sentences and by the the increasing percentage of the **Blank** sentences).

With part-of-speech tags, we observe that although using larger tag contexts can find corrections for a significantly larger number of sentences as compared to similar-sized word contexts (as indicated by the larger percentages of **Found** sentences), doing so yields overall worse MRR values. This is primarily due to the fact that with larger part-of-speech contexts the system produces more suggestions that never contain the true correction, i.e., an increasing percentage of **Missed** sentences. The most likely reason is that significantly reducing the vocabulary size by using part-of-speech tags introduces a lot of noise.

Figure 2 shows the distribution of the rank R of the true correction in the list of suggestions.[4] The figure uses a rank of 10+ to denote all ranks greater than 10 to conserve space. We observe similar trends in the figure as in Table 1 — using larger word contexts yield higher accuracies but significantly lower coverage and using larger

[3]We do not include **Blank** sentences when computing the MRR values.

[4]Note that in this figure, the bar for R = 0 includes both sentences where no ranked list was produced (**Blank**) and those where the true correction was not produced as a suggestion at all (**Missing**).

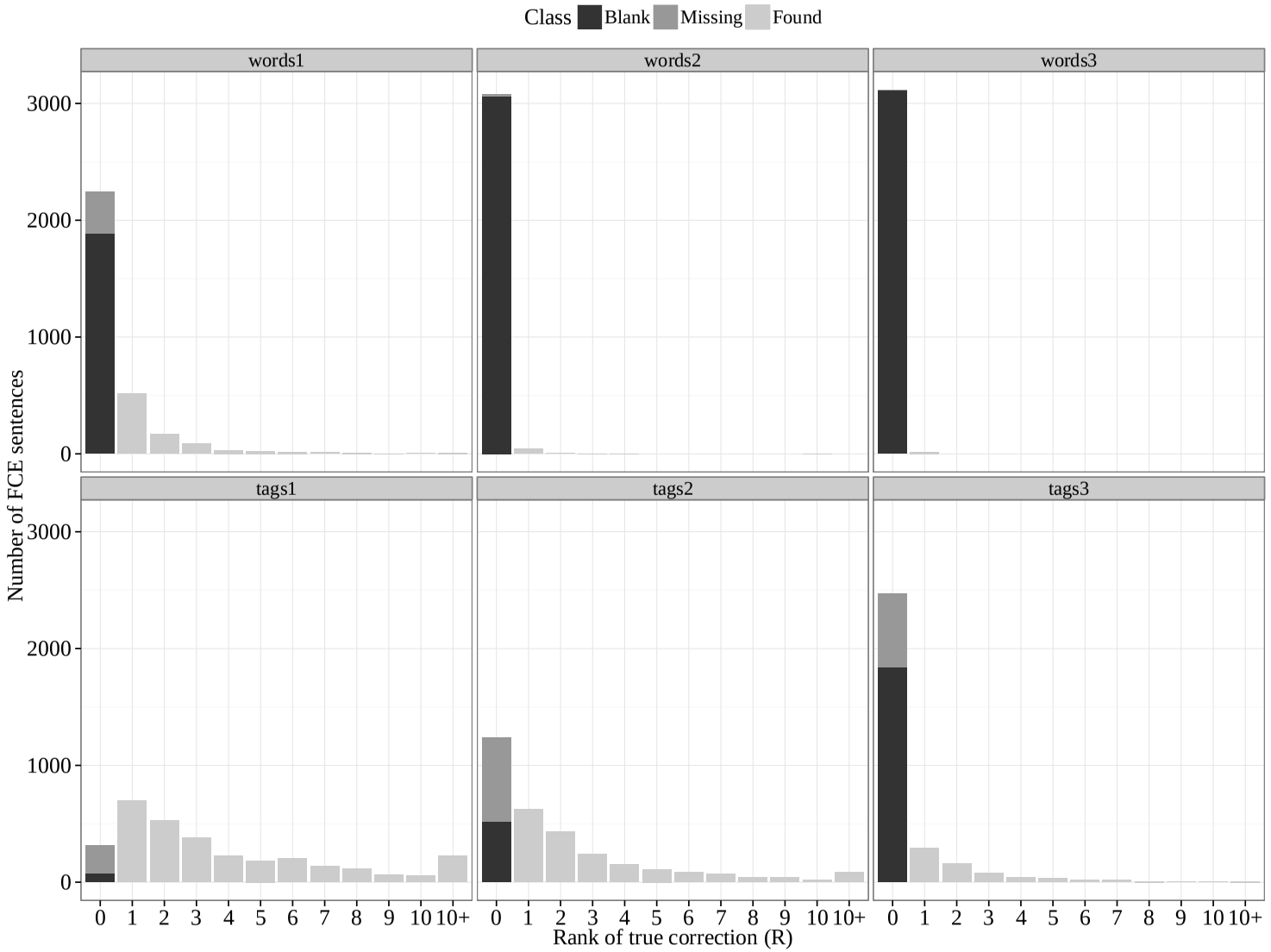

Figure 2: The distribution of the rank that the true correction has in the list of suggestions for the FCE sentences, across each context type and size used.

tag contexts yield lower accuracies and lower coverage, even though the coverage is significantly larger than that of the correspondingly sized word context.

6 Query Expansion

The results in the previous section indicate that although we could use part-of-speech tags as contexts to improve the coverage of the tool (as indicated by the number of **Found** sentences), doing so leads to a significant reduction in accuracy, as indicated by the lower MRR values.

In the field of information retrieval, a common practice is to expand the query with words similar to words in the query in order to increase the likelihood of finding documents relevant to the query (Spärck-Jones and Tait, 1984). In this section, we examine whether we can use a similar technique to improve the coverage of the tool.

We employ a simple query expansion technique for the cases where no results would otherwise be returned by the tool. For these cases, we first obtain a list of K words similar to the two words around the error from a distributional thesaurus (Lin, 1998), ranked by similarity. We then generate a list of additional queries by combining these two ranked lists of similar words. We then run each query in the list against the Wikipedia index until one of them yields results. Note that since we are using a word-based thesaurus, this expansion technique can only increase coverage when applied to the `words1` condition, i.e., single word contexts. We investigate K = 1, 2, 5, or 10 expansions for each of the context words.

Table 2 shows the a detailed breakdown of the distribution of the three classes and the MRR values with query expansion integrated into the tool for sentences where it would generally produce no output. Each row corresponds to a different value of K – the number of expansions used per context word – is varied. Note that K = 0 corresponds to the condition where query expansion is not used. From the table, we observe that using query expansion indeed seems to increase the coverage of the tool as indicated by the increasing percentage of **Found** sentences and decreasing percentage of **Blank** sentences. However, we also find that using query expansion yields worse MRR values, again because of the increasing percentage of **Missed** sentences. This represents a traditional trade-off scenario where accuracy can be traded off for an increase in coverage, depending on the desired operating characteristics.

7 Discussion and Future Work

There are several issues that merit further discussion and possibly provide future extensions to the work described in this paper.

- **Need for an extrinsic evaluation**. Although our intrinsic evaluation clearly shows that the tool has reasonably good coverage as well as accuracy on publicly available data containing preposition errors, it does not provide any evidence that the explicit feedback provided by the tool is useful to English language learners in a classroom setting. In the future, we plan to conduct a controlled study in a classroom setting that measures, for example, whether the students that see the improved feedback from the tool learn more or better than those who either see no feedback at all or those who see only implicit feedback. Biber et al. (2011) review several previously published studies on the effects of feedback on writing development in classrooms. Although the number of studies that were included in the analysis is small, some patterns did emerge. In general, students improve their writing when they receive feedback, however greater gains are made when they are presented with comments rather than direct location and correction of errors. It is unclear how students would react to a ranked list of suggestions for a particular error at a given location. An interesting finding was that L2-English students showed greater improvements in writing when they received either feedback from peers or computer-generated feedback than when they received feedback from teachers.

- **Assuming a single true correction**. Our evaluation setup assumes that the single correction provided as part of the FCE data set is the only correct preposition for a given sentence. However, it is well known in the grammatical error detection community that this is not always the case. Most usage errors such as preposition selection errors are a matter of degree rather than simple rule violations such as number agreement. As a consequence, it is common for two native English speakers

Context	K	Found	Missed	Blank	MRR
words1	0	889 (28.4%)	356 (11.4%)	1889 (60.3%)	.522
words1	1	932 (29.7%)	417 (13.3%)	1785 (57.0%)	.513
words1	2	1033 (33.0%)	550 (17.6%)	1551 (49.5%)	.493
words1	5	1118 (35.7%)	691 (22.1%)	1325 (42.3%)	.476
words1	10	1160 (37.0%)	780 (24.9%)	1194 (38.1%)	.465

Table 2: A detailed breakdown of the **Found**, **Missing** and **Blank** classes along with the Mean Reciprocal Rank (MRR) values, for different number of query expansions (K).

to have different judgments of usage. In fact, this is exactly why the tool is designed to return a ranked list of suggestions rather than a single suggestion. Therefore, it is possible that our intrinsic evaluation is underestimating the performance of the tool.

- **Practical considerations for deployment**. In this study, we used the gold standard error annotations for detecting preposition errors before querying the tool for suggestions. Such a setup allowed us to separate the problems of error detection and the generation of feedback and likely gives an upper bound on performance. Using a fully automatic error detection system will likely introduce additional noise into the pipeline, however, we believe that tuning the detection system for higher precision could mitigate that effect. Another useful idea would be to use the classifier-based approach (see §3) as a backup for the corpus-based approach for providing suggestions, i.e., using the classifier to predict the suggested corrections when no corrections can be found in the Wikipedia revisions.

- **Using other types of expansions**. In this paper, we used a very simple method of generating query expansions – a distributional thesaurus. However, in the future, it may be worth exploring other distributional similarity methods such as Brown clusters (Brown et al., 1992; Miller et al., 2004; Liang, 2005) or *word2vec* (Mikolov et al., 2013).

8 Conclusions

In this paper, we presented our work on building a proof-of-concept tool that can provide automated explicit feedback for preposition errors. We used an existing, error-annotated preposition corpus produced by mining Wikipedia revisions (Cahill et al., 2013) to not only provide a ranked list of suggestions for any given preposition error but also to produce human-authored evidence for each suggested correction. The tool can use either words or part-of-speech tags around the error as context. We evaluated the tool in terms of both accuracy and coverage and found that: (1) using larger context window sizes for words increases accuracy but reduces coverage due to sparsity (2) using part-of-speech tags leads to increased coverage compared to using words as contexts but decreases accuracy. We also experimented with query expansion for single words around the error and found that it led to an increase in coverage with only a slight decrease in accuracy; using a larger set of expansions added more noise. In general, we find that the approach of using a large error-annotated corpus to provide explicit feedback to writers performs reasonably well in terms of providing ranked lists of alternatives. It remains to be seen how useful this tool is in a practical situation.

Acknowledgments

We would like to thank Beata Beigman Klebanov, Michael Heilman, Jill Burstein, and the anonymous reviewers for their helpful comments about the paper. We also thank Ani Nenkova, Chris Callison-Burch, Lyle Ungar and their students at the University of Pennsylvania for their feedback on this work.

References

Yigal Attali. 2004. Exploring the Feedback and Revision Features of *Criterion*. Paper presented at the National Council on Measurement in Education (NCME), Educational Testing Service, Princeton, NJ.

Douglas Biber, Tatiana Nekrasova, and Brad Horn. 2011. The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis.

Research Report RR-11-05, Educational Testing Service, Princeton, NJ.

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003. Criterion online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of IAAI*, pages 3–10, Acapulco, Mexico.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust Systems for Preposition Error Correction Using Wikipedia Revisions. In *Proceedings of NAACL*, pages 507–517, Atlanta, GA, USA.

Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. The Utility of Article and Preposition Error Correction Systems for English Language Learners: Feedback and Assessment. *Language Testing*, 27(3):419–436.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical Error Correction with Alternating Structure Optimization. In *Proceedings of ACL-HLT*, pages 915–923, Portland, Oregon, USA.

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

Michael Gamon, Claudia Leacock, Chris Brockett, William B Dolan, Jianfeng Gao, Dmitriy Belenko, and Alexandre Klementiev. 2009. Using Statistical Techniques and Web Search to Correct ESL Errors. *CALICO Journal*, 26(3):491–511.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of EURALEX*, pages 105–116.

Percy Liang. 2005. Semi-supervised Learning for Natural Language. Master's thesis, Massachusetts Institute of Technology.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of ACL-COLING*, pages 768–774, Montreal, Quebec, Canada.

Anastasiya A. Lipnevich and Jeffrey K. Smith. 2008. Response to Assessment Feedback: The Effects of Grades, Praise, and Source of Information. Research Report RR-08-30, Educational Testing Service, Princeton, NJ.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name Tagging with Word Clusters and Discriminative Training. In *Proceedings of HLT-NAACL*, pages 337–342, Boston, MA, USA.

John Milton and Vivying SY Cheng. 2010. A Toolkit to Assist L2 Learners Become Independent Writers. In *Proceedings of the NAACL Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 33–41, Los Angeles, CA, USA.

John Milton. 2006. Resource-rich Web-based Feedback: Helping learners become Independent Writers. *Feedback in second language writing: Contexts and issues*, pages 123–139.

JoAnn Leah Rock. 2007. The Impact of Short-Term Use of Criterion on Writing Skills in Ninth Grade. Research Report RR-07-07, Educational Testing Service, Princeton, NJ.

Alla Rozovskaya and Dan Roth. 2010. Training Paradigms for Correcting Errors in Grammar and Usage. In *Proceedings of NAACL-HLT*, pages 154–162, Los Angeles, California.

Hongsuck Seo, Jonghoon Lee, Seokhwan Kim, Kyusong Lee, Sechun Kang, and Gary Geunbae Lee. 2012. A Meta Learning Approach to Grammatical Error Correction. In *Proceedings of ACL (short papers)*, pages 328–332, Jeju Island, Korea.

Mark D. Shermis, Jill C. Burstein, and Leonard Bliss. 2004. The Impact of Automated Essay Scoring on High Stakes Writing Assessments. In *Annual Meeting of the National Council on Measurement in Education*.

Karen Spärck-Jones and J. I. Tait. 1984. Automatic Search Term Variant Generation. *Journal of Documentation*, 40(1):50–66.

Joel R. Tetreault and Martin Chodorow. 2008. The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of COLING*, 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 ACL (short papers)*, pages 353–358, Uppsala, Sweden.

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, Edmonton, Canada.

John Truscott. 2007. The Effect of Error Correction on Learners' Ability to Write Accurately. *Journal of Second Language Writing*, 16(4):255–272.

Ellen M. Voorhees. 1999. The TREC-8 Question Answering Track Report. In *Proceedings of the Text REtrieval Conference (TREC)*, volume 99, pages 77–82.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the ACL: HLT*, pages 180–189, Portland, OR, USA.