Detecting Preposition Errors to Target Interlingual Errors in Second Language Writing

Natawut Monaikul and Barbara Di Eugenio
University of Illinois at Chicago
{monaiku1, bdieugen}@uic.edu

Abstract

Second language learners studying languages with a diverse set of prepositions often find preposition usage difficult to master, which can manifest in second language writing as preposition errors that appear to result from transfer from a native language, or *interlingual errors*. We envision a digital writing assistant for language learners and teachers that can provide targeted feedback on these errors. To address these errors, we turn to the task of preposition error detection, which remains an open problem despite the many methods that have been proposed. In this paper, we explore various classifiers, with and without neural network-based features, and fine-tuned BERT models for detecting preposition errors between verbs and their noun arguments.

1 Introduction

For learners of a second language (L2) that has complex preposition usage, choosing the correct preposition can be a challenging task. One analysis of essays written by learners of English as a second language (ESL) has found that preposition errors are the second most common error type (Leacock et al. 2010). One of the challenges in selecting the correct preposition is in the often arbitrary usage of prepositions; for example, compare English *depend **on*** with French *dépendre **de*** "depend **of/from**." Many preposition errors in L2 writing then appear to stem from the writer using a typical but erroneous translation of a preposition from a source language – a class of *interlingual errors* (James 2013).

Our ultimate goal is to build a digital writing assistant that can detect these prepositional interlingual errors, highlight the differences in usage, and provide contrastive feedback. This would be a useful tool for language learners and instructors to learn from and understand the potential sources of the errors. A necessary ability for such a system is preposition error detection (PED), a problem that has been given much attention. While many approaches have been proposed, few explicitly address the case of missing prepositions, and most consider only ESL texts.

In this work, we developed classifiers and neural network-based language models for PED between verbs and their noun arguments in ESL essays and German as a Second Language (GSL) essays. We chose to investigate this usage since it is a common site for prepositional interlingual errors (Alonso 1997; Falhasiri et al. 2011) and so that we can better localize a subset of errors of omission. We also compare these results to a fine-tuned transformer-based model using Google's multilingual BERT. In this paper, we show that (1) a random forest classifier with a modest set of features can perform with 57.24% precision at 32.16% recall on ESL essays, which outperforms our fine-tuned BERT models (2) enhancing our classifiers with our neural network language model improves their performance, and (3) our enhanced random forest model can detect 46.5% of missing prepositions after verbs in ESL essays.

2 Related Work

The task of PED has been popularly approached with maximum entropy classifiers, which can be trained on native English text to serve as a language model and then be used to evaluate the appropriateness of a preposition in a piece of ESL text given its context (Chodorow, Tetreault, and Han 2007; De Felice and Pulman 2008); alternatively, the classifier can be trained directly on error-annotated ESL text, instead modeling whether or not a preposition has been used correctly (Han et al. 2010).

These high-precision systems, however, only address the cases in which the preposition actually appears in the text, leaving out the case in which a preposition is missing. One system that handles this case uses heuristic rules on sequences of part-of-speech tags to determine potential sites of omitted prepositions (Gamon et al. 2008).

Furthermore, these methods only report results on ESL essays. While there has been work in grammatical error detection in GSL writing, the methods generally involve parsing using specialized grammars and lexicons, targeting mainly word order and agreement errors (Heift and Nicholson 2001). In this work, we present methods evaluated on both ESL and GSL writing. We also note that while systems such as Grammarly (https://www.grammarly.com/) perform PED, these proprietary systems cannot be easily integrated with another feedback system.

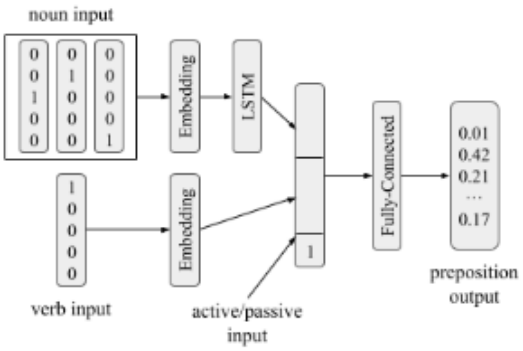

Figure 1: Architecture of our proposed neural network to capture preposition usage in their contexts

3 Methods

In both English and German, a noun argument of a verb can be introduced by a preposition or without a preposition (e.g., as a direct object), which we treat as a *null* preposition. These relationships are encompassed in *dependency trees* (Jurafsky and Martin 2009), in which words are labeled for their grammatical relation to other words in the sentence. We posit that the verb and noun argument used are major contributors to selecting the intermediate preposition, and so we propose approaches to PED that use classifiers with features meant to capture these co-occurrences.

3.1 Features

We developed a set of numerical features to represent the appropriateness of a preposition given its context: the verb and the noun argument it introduces. We tabulated frequencies of prepositions with each verb and noun argument from a corpus of 100k movie reviews in English from the Internet Movie Database (IMDb) (Maas et al. 2011) and from a corpus of 1.2M sentences from randomly chosen websites in German (Goldhahn, Eckart, and Quasthoff 2012). We chose these corpora for their comparable level of formality and use of first- and second-person pronouns as the ESL and GSL datasets we used (described in Section 4.1).

From the tabulated frequencies, we calculated eight features that capture the likelihood of the preposition appearing with the given verb and with the given noun, and the likelihood of the given verb being used *transitively* (with a direct object). We also include a feature to indicate whether or not the verb is in the passive voice (e.g., *compose a song* vs. *is composed **of** wood*). These features were then used in logistic regression (LR) and random forest (RF) classifiers.

3.2 Neural Network Language Model

We also developed a neural network language model for the purpose of learning to predict preposition usage given a verb and noun argument context from each corpus, potentially capturing a more complex relationship between the preposition, verb, and noun.

The architecture of our network is given in Figure 1. For the noun argument, three noun arguments of the verb are given in sequence – the previous, target, and next noun arguments. This is intended to account for cases when the order of noun arguments affects prepositions usage (e.g., *write you a letter* vs. *write a letter **to** you*) and for cases of multiword expressions (e.g., *take part **in** something*). Separate embedding layers are used for verbs and nouns because the layers are meant to learn and represent how prepositions co-occur with verbs and nouns, the usage of which can be independent. The final output is a probability distribution over a set number of prepositions. This output can then be used as additional features for the LR and RF classifiers by taking the predicted value of the preposition used by the writer and the greatest value in the output. We denote these hybrid classifiers as LR+NN and RF+NN.

3.3 BERT Language Model

Another language model we explored is the Bidirectional Encoder Representations from Transformers (BERT), recently released by Google (Devlin et al. 2018). BERT uses a transformer model (stacked encoders and decoders with self-attention) that has been trained on a word prediction task and a next-sentence prediction task to learn contextualized word- and sentence-level embeddings. Pre-trained BERT models have been fine-tuned for many popular NLP tasks, producing state-of-the-art results.

We therefore fine-tuned a pre-trained BERT model for PED. Given an ESL or GSL sentence and a preposition used in the sentence, we would like the model to predict whether or not there is a preposition error. We give the model two inputs: the sentence with the target location masked, and the preposition used (which can be `NONE`). The model outputs a vector representation for each input token, as well as for a special `CLS` token meant for classification tasks. This `CLS` token representation can then be sent through a final fully-connected layer for our desired prediction.

4 Evaluation

4.1 Dataset

To evaluate our proposed approaches on ESL text, we used a dataset of 2,481 error-annotated essays written by Cambridge ESOL First Certificate in English (FCE) examinees from a variety of language backgrounds (Yannakoudakis, Briscoe, and Medlock 2011). These annotations include preposition errors (missing, denoted "MT"; extraneous, denoted "UT"; and incorrect choice, denoted "RT").

For comparable GSL text, we used the MERLIN corpus, which contains 1,033 L2 German essays from The European Language Certificates (TELC) exams (Boyd et al. 2014). The examinees also come from a variety of language backgrounds, and the corpus has also been annotated for errors. Similar to analyses of ESL corpora, preposition errors were the third most frequently-annotated grammatical error in these GSL essays.

4.2 Procedure

The English language IMDb corpus and the German language web corpus were parsed using *spaCy* (https://spacy.io/), which provides dependency parsing models in several

	ESL (FCE)				
	LR	**RF**	**LR+NN**	**RF+NN**	**BERT**
Precision	16.31	57.24	17.79	62.58	76.94
Recall	55.28	32.16	62.47	30.82	14.01
F1-Score	25.19	41.18	27.69	**41.30**	23.70
	GSL (MERLIN)				
	LR	**RF**	**LR+NN**	**RF+NN**	**BERT**
Precision	11.93	36.70	13.99	43.44	14.37
Recall	24.82	13.88	28.33	13.83	26.16
F1-Score	16.11	20.14	18.73	**20.98**	18.55

Table 1: Precision, recall, and F1-score (in %) for all classifiers on the ESL and GSL datasets (highest F1-score bolded)

languages. The resulting dependency trees were then used to extract the data points needed for each of our models. This resulted in about 1.7M data points, of which 44% had a non-null preposition, across 13k distinct verbs in the English corpus, and 2M data points, of which 48% had a preposition, across 35k distinct verbs in the German corpus.

The *spaCy* dependency parser was also used on the FCE and MERLIN datasets to extract similar data points. Since both datasets are error-annotated, each data point was labeled with whether or not there was a preposition error in the extracted data point. In total, 40,870 data points were collected from the FCE dataset, of which 2,368 contained a preposition error (535 MT, 413 UT, 1,420 RT), capturing 40.4% of all preposition errors, and 13,611 data points were collected from the MERLIN dataset, of which 434 contained a preposition error (113 MT, 94 UT, 227 RT), capturing 49.5% of all preposition errors.

The LR and RF classifiers were implemented using the *scikit-learn* toolkit for machine learning (Pedregosa et al. 2011) with default parameters. Our neural network language models were implemented using the *keras* toolkit (Chollet and others 2015), with embedding and LSTM layers of size 300 with default activations and fully-connected layers of size 300 with ReLU activations.

The English and German language data points were used to train separate language models to output a softmax distribution over the 30 most frequent prepositions in the datasets. 90% of the data points with no preposition were randomly excluded to alleviate major skewing. Each model was trained for 10 epochs with a batch size of 1,024. The English-based model reached 69.2% accuracy in predicting the correct prepositions in its training data, and the German-based model achieved 65.5%.

The *BERT-Base, Multilingual Cased* pre-trained model (retrieved from https://github.com/google-research/bert) was used for our fine-tuning, one for each dataset. Because of the computing power needed for BERT, we used Google Cloud Platform's TPU for fine-tuning.

4.3 Results

Five-fold cross-validation was used for evaluation. The performance of each classifier is reported as the average precision, recall, and F1-score across the five folds in Table 1.

In both the ESL and GSL datasets, RF+NN outperformed the other classifiers by F1-score. As a comparison, we may look to the results of the HOO 2012 shared task on PED (Dale, Anisimoff, and Narroway 2012), which also used the FCE dataset. The submission with the highest F1-score achieved 41.47 (Rozovskaya, Sammons, and Roth 2012) – comparable to our RF+NN F1-score – and the submission with the highest precision achieved 56.99% precision and 22.46% recall (Dahlmeier, Ng, and Ng 2012). We note, however, that the shared task targeted all preposition errors, while our approach focuses on a subset of those errors.

The addition of the two neural network-based features slightly improved the F1-scores of the LR and RF classifiers for both datasets. The difference was significant for the LR classifiers (via paired t-tests, $p < 0.05$), but not significant for the RF classifiers. One limitation of the neural network-based features is the way in which the networks were trained. For each given verb and noun argument, only one preposition is given as the correct preposition; however, there are many instances in which more than one preposition can be used and still be grammatically correct, whether it changes the meaning of the phrase (e.g., *go **to** the store* vs. *go **in** the store*) or not (e.g., *wait a minute* vs. *wait **for** a minute*). A future direction would be to treat this preposition prediction task as a multilabel classification problem, instead of single label. Another reason could be not adequately capturing the information in the output of the networks; we also plan to explore more sophisticated measures pertaining to probability distributions.

We also found that the RF models (with and without NN) outperformed the BERT fine-tuned models. It is possible that BERT was unable to produce effective sentence-level representations in this task because many of the sentences, since they are written by non-native speakers of the language, contain misspellings and grammatical errors, while BERT was mainly pre-trained on native English/German texts. Another consideration (which can apply to all of the classifiers we explored) is that the dataset is extremely unbalanced: in the FCE dataset, only 5.8% of the extracted data points contained a preposition error, and in the MERLIN dataset, only 3.2%. Given the number of parameters in the BERT model, this may have affected the confidence with which the fine-tuned models could label errors. While an effort can be made to balance the dataset via undersampling or oversampling, this would not be representative of the input that a writing feedback system would receive – an essay in which most of the prepositions used (and not used) are correct.

In general, our classifiers performed better on the ESL dataset than on the GSL dataset. This could be due to the size difference between the two datasets or the training differences between the English and German parsers. While the accuracy of each parser is reported by *spaCy*, this accuracy cannot reflect the more difficult task of parsing L2 texts.

Finally, since part of our rationale for choosing to target prepositions between verbs and their noun arguments was to better locate missing prepositions, we also recorded the percentage of MT errors our classifiers could capture. Across the five folds, RF+NN correctly marked 46.5% and 15.3% of MT errors on average in the ESL and GSL datasets, respectively. We compare the ESL result again to the highest pre-

cision HOO submission (since results split by error type are explicitly reported), which captures 21.05% of MT errors (Dahlmeier, Ng, and Ng 2012). However, since we focused on a subset of the preposition errors in the FCE dataset, which covers 37% of MT errors, our RF+NN actually captures about 17.3% of all MT errors in the FCE dataset. While our method can handle a comparably reasonable percentage of errors of omission, there is still much room for improvement. We plan to target other similar potential sites of missing prepositions, such as between adjectives and nouns (*interested **in** something*) and between two nouns (*millions **of** people*), in our future work.

5 Conclusion

We envision an assistive tool for second language learners and language instructors that can locate potential interlingual errors on prepositions and provide contrastive feedback. To this end, we developed and tested traditional and neural network-based classifiers, including fine-tuned BERT models, that use information from English and German texts to predict preposition errors between verbs and their noun arguments in ESL and GSL essays, respectively. We showed that a random forest classifier with a simple set of features can perform competitively in preposition error detection, improved by adding a neural language model, at a common site for omitted prepositions. We also found that while BERT has achieved state-of-the-art results in a variety of downstream NLP tasks, it may not be as powerful on non-native texts.

6 Acknowledgements

This work is partly supported by NSF award IIS 1705058 and by a University Scholar award to Barbara Di Eugenio from the University of Illinois at Chicago (UIC).

References

Alonso, M. R. A. 1997. Language transfer: Interlingual errors in Spanish students of English as a foreign language. *Revista Alicantina de Estudios Ingleses* (10):7–14.

Boyd, A.; Hana, J.; Nicolas, L.; Meurers, D.; Wisniewski, K.; Abel, A.; Schöne, K.; Štindlová, B.; and Vettori, C. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Chodorow, M.; Tetreault, J. R.; and Han, N.-R. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, 25–30.

Chollet, F., et al. 2015. Keras. https://keras.io.

Dahlmeier, D.; Ng, H. T.; and Ng, E. J. F. 2012. NUS at the HOO 2012 shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 216–224.

Dale, R.; Anisimoff, I.; and Narroway, G. 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*, 54–62.

De Felice, R., and Pulman, S. G. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 169–176.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Falhasiri, M.; Tavakoli, M.; Hasiri, F.; and Mohammadzadeh, A. R. 2011. The effectiveness of explicit and implicit corrective feedback on interlingual and intralingual errors: A case of error analysis of students' compositions. *English Language Teaching* 4(3):251–264.

Gamon, M.; Gao, J.; Brockett, C.; Klementiev, A.; Dolan, W. B.; Belenko, D.; and Vanderwende, L. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing*.

Goldhahn, D.; Eckart, T.; and Quasthoff, U. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 759–765.

Han, N.-R.; Tetreault, J. R.; Lee, S.-H.; and Ha, J.-Y. 2010. Using an error-annotated learner corpus to develop an ESL/EFL error correction system. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Heift, T., and Nicholson, D. 2001. Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education* 12:310–325.

James, C. 2013. *Errors in Language Learning and Use: Exploring Error Analysis*. Routledge.

Jurafsky, D., and Martin, J. H. 2009. *Speech and Language Processing*, volume 2. Prentice Hall.

Leacock, C.; Chodorow, M.; Gamon, M.; and Tetreault, J. 2010. Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies* 3(1):1–134.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Rozovskaya, A.; Sammons, M.; and Roth, D. 2012. The UI system in the HOO 2012 shared task on error correction.

Yannakoudakis, H.; Briscoe, T.; and Medlock, B. 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*, 180–189.