TABLE II
RESULTS FOR EXPERIMENT 1, BASED ON VARIOUS NUMBER OF BERT PREDICTIONS, N

top-N	accuracy top@1	accuracy top@N	P(top@1|top@N)
10	48.77%	49.64%	98.25%
50	64.34%	66.80%	96.32%
100	68.34%	71.71%	95.30%
200	71.08%	75.52%	94.13%
300	72.14%	77.20%	93.45%
400	72.92%	78.17%	93.28%
500	**73.25%**	78.84%	92.91%

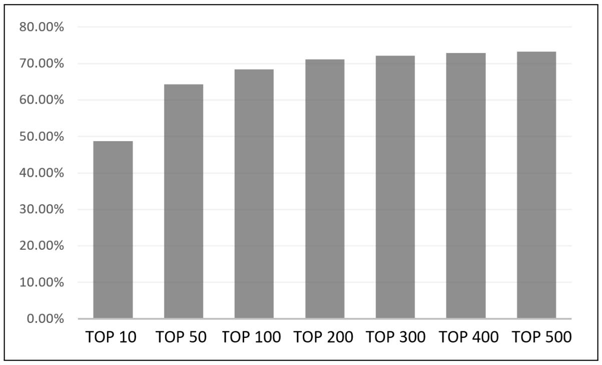

Fig. 1. Final accuracy changes when increasing N

The metric P(top@1|top@N) showed a clear decreasing trend. Combining the accuracy top@1 and the P(top@1|top@N), we can expect that the accuracy top@1 may converge.

B. Experiment 2: Applying BERT after edit distance

As stated earlier, in this experiment we first generate a list of similar words within the K edit distance of the misspelled token, then apply BERT to rank the candidate corrections. BERT is capable of computing a softmax probability for each candidate word within the provided context. Thus, the word with the highest probability is chosen as the final prediction.

As mentioned in the previous section, edit distance can also be used as a criteria to filter lexical entities from an existing corpus. Without the corpus restriction, many non-word contents are generated through modifying the misspelled token, which leaves BERT's vocabulary as the only criteria to restrict real-word selection. When a particular corpus is applied prior to BERT's selection, it is feasible for common words or domain specific vocabularies to be captured to further improve the accuracy of the corrections.

Note that in experiment 1, a threshold is not required for edit distance to realize the selection process as the minimal value is always preferred. In this experiment, as the edit distance is applied as the first round selection criteria, a threshold is required since it is impossible to generate candidates with an infinitely large K (one could keep adding letters to a word). Such restriction of K can affect the accuracy of the corrections, since the labels can be out-of-vocabulary words which are filtered out before the subsequent steps.

In order to compare how edit distance and corpus can jointly affect the candidate selections, we control the example sentences sampled from the dataset to eliminate out-of-vocabulary words. Three subsets of sentences are created based on the following mechanism:

- sample the sentence if the label is within K edit distance of the misspelled token.
- sample the sentence if the label is also in the corpus.
- sample the sentence if the label is within K edit distance of the misspelled token and that the word is in the corpus.

In this particular experiment, an edit distance of K $\leq$ 2 and the Brown corpus are used for all sub-tasks. Among the 2075 testing cases, 1934 of them have K $\leq$ 2. If we increased the K to 3 or 4, the coverage barely grows but the cost of computation can grow exponentially. TABLE III shows some examples of words and the number of similar spellings with different edit distance:

TABLE III
NUMBER OF SIMILAR WORDS WITH DIFFERENT EDIT DISTANCE (ED)

Word	ED = 1	ED = 2	ED = 3	ED = 4
study	4	35	427	3148
annoying	1	2	42	480
adventurous	1	3	7	13

The top@1 accuracy for each restricted subsets, as described below, are compared with a comparison group with no restrictions on either edit distance or corpus. The results are compared between the following groups:

- **no restriction**: no sub-sampling on the dataset.
- **ED2 only**: subset containing examples if the label is within 2 edit distance of the misspelled token.
- **corpus only**: subset containing examples if the label is also in the Brown corpus.
- **ED2 + corpus**: subset containing examples if the label is within 2 edit distance of the misspelled token and that the word is in the Brown corpus.

TABLE IV shows the result for each group.

TABLE IV
EXPERIMENT 2 RESULT

	Total Cases	Correct Cases	Accuracy
No Restriction	2075	1491	71.86%
ED2 only	1934	1491	77.09%
corpus only	1885	1491	79.10%
ED2 + corpus	1756	1491	84.91%

VI. DISCUSSION

A. BERT can be used for misspelling correction

In this experiment, the BERT model was not specially trained or fine-tuned. It was not designed for misspelling correction task but still provided acceptable result in both experiments. After applying the edit distance algorithm with the BERT model, 73.25% of the misspellings were corrected successfully in experiment 1. If we exclude the out-of-vocabulary words and the errors with large edit distance in experiment 2, the accuracy can reach 84.91%. The experiments showed using BERT to do misspelling correction was possible if we treated the misspelling correction task as masked word prediction task.

BERT can always be fine-tuned with new training corpus. If the BERT model can be trained on some texts related to the topic of the chosen data set (for instance, some exam scripts for the same questions without errors), there might still be room for further improvement.

Other BERT-like models trained on larger corpus are likely to perform better due to the larger vocabulary size and richer topics of the texts. However, models like Generative Pre-trained Transformer 3 (GPT-3) [24] require enormous computation power to do inferences, which are less practical in the use case of misspelling correction.

B. Analysis on Part of Speech of the misspellings

Among all the testing cases, some of the errors seem easy to be identified and corrected by human — here are 2 examples:

example 1: "*Another point which I think was ennoying (correction: annoying) was the concert hall*"

example 2: "*There, I saw the exibitions (correction: exhibitions) and admired the building itself*"

Both corrections are only 1 edit distance away from the misspellings. However, in experiment 1, even if we increase the top-N to 500, BERT and the edit distance algorithm still failed to find the correct answer. BERT provided a list of words that matched the Part of Speech (POS) and made sense in the given context. It is possible that if we masked a word with a certain POS, BERT can be confused due to the lack of context.

In experiment 2, the misspelling "exibitions" in example 2 was corrected as "exhibition" which was the single form of the expected correction. According to the context, both forms are acceptable in terms of grammar, but it was counted as a failed case in the evaluation. For certain types of POS, we might need to add extra limitations during the evaluation to eliminate the cases like example 2.

C. BERT can be tolerant of a certain amount of error in the context

All the testing cases were written by non-native speakers. Many of sentences contained multiple errors. Not only the masked word itself was a misspelling, but there were other types of errors in the sentence as well. BERT showed great robustness in such situations. To further prove this observation, we even tested some informal sentences from social media and BERT still worked properly. Here are some examples:

example 1: "*whos gonna tell (masked) my brother that the mf corona is here already and he needs to stay the [...]*[1] *out from society for awhile*"

example 2: "*seal the whole area let the people be there anyone wants to come in give them a 12hour deadline after that set up a perimeter around the bagh and let them be one (masked) gets out or in till corona is over*"

The second example was actually a paragraph consisted of multiple sentences. The punctuation was missing, and the spellings were chaotic. Even in such situation, BERT still was able to predict the masked word successfully.

VII. Future Work

A. Misspellings with larger edit distance

In experiment 2, only similar words with edit distance under 2 were selected as the candidate words due to the balance between accuracy and computation cost. However, in reality, there are spelling errors with a larger edit distance. For example, sometimes people create abbreviations for the long words based on the pronunciations or the spellings. The method used in experiment 2 may not be capable of correcting such errors with the current algorithms. Future work is required to solve this problem.

B. Other types of errors

If misspelling correction can be converted to a masked word prediction task, what about correcting other types of errors? For example, some syntax error correction task can also be viewed as a masked prediction task. Some preliminary experiments were also conducted that considered sentences such as:

"*The book is going to change the way people thinking.*"

The example above contained a syntax error. Either the verb "*thinking*" was in a wrong form, or, less likely, the words "will be" are missing.

The pre-trained BERT model was very good at masked word prediction as long as the expected output was one word. With this underlying idea, BERT could become a universal model that is capable of correcting different types of errors. The CLC FCE dataset contains different types of error. Some of them are strictly in one-to-one format (the error was one word and the correction was also one word). It would be interesting to compare the results across different types of error.

C. BERT for both misspelling detection and correction

In the experiments described in this paper, the misspellings were labeled. Therefore BERT did not need to detect the misspelling before fixing it. In real-world application, misspelling detection and correction are always bounded together. It would be more useful if BERT could also be used for misspelling detection.

One approach that allows BERT to detect anomalies could be: scan each of the word in a given sentence and try to find anomaly based on the BERT prediction probability and the edit distance. The metrics need to be carefully designed and tested.

For every single masked word prediction task, BERT would compute the attention scores for every word in the input

[1] The word is not displayed in the paper due to its offensive nature

sentence. Thus the speed of scanning every word in a sentence is similar to the speed of doing one masked work prediction task.

VIII. Conclusion

This paper explored the possibility of using the pre-trained BERT model for misspelling correction. We utilized the masked word prediction capability of BERT to do misspelling correction. The test result showed the pre-trained BERT model was capable of fixing spelling errors with the help of the edit distance algorithm.

During our experiments, we did not train or fine-tune the BERT model. Although the accuracy was not close to the ngram model [8] or the RNN model [12], the pre-trained BERT model proved its potential in the misspelling correction task.

Acknowledgment

We would like to thank Baijian Yang for his valuable and helpful feedback during our research progress and Shinhye Yun for her input in the early stages of this work.

References

[1] Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. ”Allennlp: A deep semantic natural language processing platform.” arXiv preprint arXiv:1803.07640 (2018).

[2] Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. ”Deep contextualized word representations.” arXiv preprint arXiv:1802.05365 (2018).

[3] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

[4] D. Jurafsky and J. H. Martin, “Spelling correction and the noisy channel,” in Speech and language processing, Draft of October 2, 2019. [Online]. Available: https://web.stanford.edu/ jurafsky/ slp3/26.pdf

[5] Church, Kenneth W., and William A. Gale. ”Probability scoring for spelling correction.” Statistics and Computing 1, no. 2 (1991): 93-103.

[6] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. ”Attention is all you need.” In Advances in neural information processing systems, pp. 5998-6008. 2017.

[7] Mays, Eric, Fred J. Damerau, and Robert L. Mercer. ”Context based spelling correction.” Information Processing & Management 27, no. 5 (1991): 517-522.

[8] Carlson, Andrew, and Ian Fette. ”Memory-based context-sensitive spelling correction at web scale.” In Sixth International Conference on Machine Learning and Applications (ICMLA 2007), pp. 166-171. IEEE, 2007.

[9] GNU Aspell. (1998). [Online]. Available: http://aspell.net/

[10] Gupta, Prabhakar. ”A context-sensitive real-time Spell Checker with language adaptability.” In 2020 IEEE 14th International Conference on Semantic Computing (ICSC), pp. 116-122. IEEE, 2020.

[11] Hunspell. [Online]. Available: http://hunspell.github.io/.

[12] Li, Hao, Yang Wang, Xinyu Liu, Zhichao Sheng, and Si Wei. ”Spelling error correction using a nested rnn model and pseudo training data.” arXiv preprint arXiv:1811.00238 (2018).

[13] Sakaguchi, Keisuke, Kevin Duh, Matt Post, and Benjamin Van Durme. ”Robsut wrod reocginiton via semi-character recurrent neural network.” In Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[14] Kim, Yoon, Yacine Jernite, David Sontag, and Alexander M. Rush. ”Character-aware neural language models.” In Thirtieth AAAI conference on artificial intelligence. 2016.

[15] Ge, Tao, Furu Wei, and Ming Zhou. ”Reaching human-level performance in automatic grammatical error correction: An empirical study.” arXiv preprint arXiv:1807.01270 (2018).

[16] Didenko, Bohdan, and Julia Shaptala. ”Multi-headed Architecture Based on BERT for Grammatical Errors Correction.” In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 246-251. 2019.

[17] Yannakoudakis, Helen, Ted Briscoe, and Ben Medlock. ”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, pp. 180-189. 2011.

[18] Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. ” O’Reilly Media, Inc.”, 2009.

[19] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. ”Transformers: State-of-the-art natural language processing.” arXiv preprint arXiv:1910.03771 (2019).

[20] Taylor, Julia M., and Victor Raskin. ”Understanding the unknown: Unattested input processing in natural language.” In 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), pp. 94-101. IEEE, 2011.

[21] Jing, Xiaonan, Penghao Wang, and Julia M. Rayz. ”Discovering Attribute-Specific Features From Online Reviews: What Is the Gap Between Automated Tools and Human Cognition?.” International Journal of Software Science and Computational Intelligence (IJSSCI) 10, no. 2 (2018): 1-24.

[22] Damerau, Fred J. ”A technique for computer detection and correction of spelling errors.” Communications of the ACM 7, no. 3 (1964): 171-176.

[23] Levenshtein, Vladimir I. ”Binary codes capable of correcting deletions, insertions, and reversals.” In Soviet physics doklady, vol. 10, no. 8, pp. 707-710. 1966.

[24] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. ”Language models are few-shot learners.” arXiv preprint arXiv:2005.14165 (2020).