# Grammatical Error Correction Using Pseudo Learner Corpus Considering Error Tendency of Learners

**Yujin Takahashi, Satoru Katsumata*[*] and Mamoru Komachi**
Tokyo Metropolitan University
takahashi-yujin@ed.tmu.ac.jp, satoru.katsumata@retrieva.jp
komachi@tmu.ac.jp

## Abstract

Recently, several studies have focused on improving the performance of grammatical error correction (GEC) tasks using pseudo data. However, a large amount of pseudo data are required to train an accurate GEC model. To address the limitations of language and computational resources, we assume that introducing pseudo errors into sentences similar to those written by the language learners is more efficient, rather than incorporating random pseudo errors into monolingual data. In this regard, we study the effect of pseudo data on GEC task performance using two approaches. First, we extract sentences that are similar to the learners' sentences from monolingual data. Second, we generate realistic pseudo errors by considering error types that learners often make. Based on our comparative results, we observe that $F_{0.5}$ scores for the Russian GEC task are significantly improved.

## 1 Introduction

Recently, several studies have proposed models to solve grammatical error correction (GEC) task as an application of writing support for language learners of various languages, such as English or Russian. A standard approach to improve GEC models is to incorporate pseudo errors into large monolingual datasets for pre-training. In particular, previous works achieved state-of-the-art performance by pre-training the model using pseudo data with a subsequent fine-tuning of the pre-trained model using a learner corpus (Zhao et al., 2019; Kiyono et al., 2019; Grundkiewicz et al., 2019; Náplava and Straka, 2019; Grundkiewicz and Junczys-Dowmunt, 2019).

Considering the aforementioned approach, several methods have been proposed for the generation of pseudo data for pre-training a GEC model. In theory, it is possible to include all types of errors in a dataset via random error generation. However, considering the limitations of computational resources required to train a GEC model using large pseudo datasets, there is a need to generate pseudo datasets with only realistic errors.

Thus, in this study, we generate pseudo data to train GEC models considering the types of errors made by language learners and study the effect of this realistic pseudo training data. First, we extract sentences similar to the training data from monolingual datasets to generate pseudo data for pre-training. Second, we analyze the error tendency of learners and add pseudo errors considering the errors learners tend to make in English and Russian languages. Through experiments, we show that the proposed pseudo data generation method improves the $F_{0.5}$ scores of the GEC model.

In summary, the primary contributions of this study are as follows:

- We confirm that selecting training data similar to the learners' corpus instead of using randomly selected monolingual data improves the performance of the GEC model.

- We show the effect of realistic pseudo errors by considering the types of errors typically made by language learners for the Russian GEC task.

## 2 Related Works

Pseudo data have been generated for GEC tasks in several previous works. Zhao et al. (2019) generated pseudo data by adding randomly generated pseudo errors, in an error-free sentence. In particular, in this approach, randomly selected words were replaced or deleted from a large monolingual dataset. In addition, a random word was inserted into sentences, and words in a sentence

[*] Currently at Retrieva, Inc.

*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 27–32
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics

| En (CoNLL 2013) | | Ru (RULEC-GEC dev) | |
|---|---|---|---|
| Error type | Ratio (%) | Error type | Ratio (%) |
| Art./Det. | 19.9 | Spelling | 22.8 |
| Collocation/Idiom | 12.5 | Insert | 13.2 |
| Noun number | 11.4 | Noun case | 10.2 |
| Preposition | 8.98 | Replace | 9.99 |
| Word form | 6.56 | Delete | 9.58 |

Table 1: Comparison of error statistics between English and Russian learner corpora (Development Data).

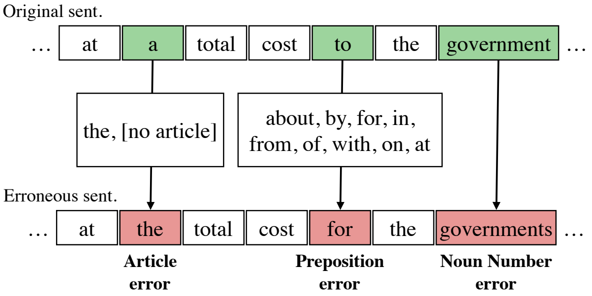
Original sent.
… at a total cost to the government …
the, [no article]
about, by, for, in, from, of, with, on, at
Erroneous sent.
… at the total cost for the governments …
Article error
Preposition error
Noun Number error

Figure 1: Example of pseudo error generation.

were swapped around. A similar approach was proposed by Kiyono et al. (2019), where an original word is masked or retained to generate pseudo data for pre-training. However, both of these methods generate errors that are not similar to the real errors made by language learners. The data in Table 1 indicates that English language learners tend to make errors related to article and word choice, while Russian language learners often make errors related to spelling, insertions, and noun inflections. In our study, we use these error tendencies to generate realistic errors to develop pre-training datasets for GEC tasks in those languages.

Furthermore, Grundkiewicz et al. (2019) generated realistic pseudo data by building a confusion set based on an unsupervised spellchecker to restrict word replacements made by learners in the resulting dataset. They used the conditional probability $P(cor|err)$ based on the spellchecker distribution; however, it is not the same as $P(err|cor)$, nor does it include error types other than spelling errors. Conversely, in our work, we approximate $P(err|cor)$ using a uniform distribution for the set of candidates for a correct word. This uniform distribution is developed using prior knowledge of error types instead of that obtained from a spellchecker. Thus, our generated pseudo data contains comparatively more realistic pseudo errors. Kasewa et al. (2018) determined the distribution of the pseudo error generation model $P(err|cor)$ from parallel data obtained using a grammatical error detection task.

Moreover, Grundkiewicz and Junczys-Dowmunt (2019) developed a confusion set that retained out-of-vocabulary words and preserved consistent letter casing. However, using this approach, unrealistic errors might be included in the pseudo data because it primarily considers the surface of words. Further, Náplava and Straka (2019) conducted a GEC experiment in multiple languages, such as English, Russian, German, and Czech, and proposed a pseudo error generation model for Czech, considering errors in diacritics. In the present study, we incorporate the most common error types in monolingual data based on language-specific prior knowledge to obtain development data.

## 3 Method for Pseudo Data Generation

First, we describe the method for pseudo data generation that considers learner error types. Subsequently, we use the generated pseudo data for pre-training a GEC model.

In this study, we combine the proposed method of pseudo data generation with previous methods. In particular, we incorporate the basic random approach (deletion, insertion, swapping) in our approach, as well as the more recent sophisticated approach proposed by Grundkiewicz et al. (2019) (character level perturb, confusion set based on an unsupervised spellchecker).

### 3.1 Data Selection

We assume that the sentences, where errors of the learners' error types are added, should be similar to that of the learners' sentences themselves. Thus, we used a data selection method (Moore and Lewis, 2010), where an N-gram language model (LM) is used to score input sentences. This method creates a generic LM $N$ and targets LM $I$ sets for the generic and target domains, respectively. Subsequently, the entropy $H$ is calculated for the sentence $s$ in monolingual data from these LM sets ($\mathrm{LM}_{\mathrm{model}} \in \{I, N\}$). Finally, the entropy difference (Equation 1) for the sentence is calculated. Data selection is then performed based on the similarity to the target domain in descending order of the assigned score.

$$\text{score}(s) = H(s; N) - H(s; I) \qquad (1)$$

$$H(s; \text{LM}_{\text{model}}) = -\frac{1}{|s|} \log P_{\text{LM}_{\text{model}}}(s)$$

where $|s|$ indicates the sentence length, $P_{\text{LM}_{\text{model}}}(s)$ indicates the probability estimated by the $\text{LM}_{\text{model}}$ for sentence $s$.

In this study, for each sentence in the monolingual data, the entropy difference is calculated between the LM trained on monolingual data and that trained on the data in the target domain. Subsequently, sentences are extracted according to the LM scores for pre-training data.

### 3.2 Error Types

Figure 1 shows an example of pseudo error generation according to the most common error types in learners' corpora. As an example of preposition errors, we limit the confusion set by defining the pseudo error generation model as $P(err|cor =$ "to") where $err \in \{$about, by, for, from, in, of, with, on, at$\}$. The pseudo error is generated using a uniform distribution for the pseudo error generation model $P(err|cor)$.

**English.** As listed in Table 1, the common error types in English are those related to article/determiner, collocation/idiom, noun number, preposition, and word form. Thus, for English, we consider each error type as follows:

- For article/determiner errors, the set of replacement candidates is the entire vocabulary in the random baseline. However, we limit the set of replacement candidates to other articles and determiners only. This set contains an entry of "no article" as well (i.e., deletion).

- For noun number errors, the error can be generated by swapping the singular or plural form of a noun with the plural or singular form, respectively.

- For preposition errors, we define a candidate set as the top 10 most frequently used prepositions (Bryant and Briscoe, 2018). We only replace the preposition with one from the candidate sets.

- For word form errors, we define a candidate set for replacement using `word_forms` [1].

| Lang. | Dataset | Corpus | Sent. |
|---|---|---|---|
| English | One Billion Corpus | mono | 10M |
| | Lang-8 + NUCLE | para | 134K |
| Russian | Russian News Crawl | mono | 10M |
| | Lang-8 + RULEC-GEC | para | 54K |

Table 2: Data statistics.

We did not consider collocation and idiom errors in our study because defining a candidate set for those error types is challenging.

**Russian.** For the Russian language, we consider replacement and spelling errors as per the previously proposed methods (i.e., random and unsupervised spellchecker). For noun case errors, we define a candidate set for replacement using a dictionary. When the target word is a noun and is included in the dictionary, the candidates for replacement consist of the inflected patterns specified in the dictionary.

## 4 Experiments

### 4.1 Data

Table 2 lists the details of monolingual and parallel data used for training in our study. As training data, we used Lang-8 (Mizumoto et al., 2012) and NUS Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013) for English, while we used Lang-8 and Russian Learner Corpus of Academic Writing-GEC (RULEC-GEC) (Rozovskaya and Roth, 2019) for Russian. As pre-training data (i.e., pseudo data), we used One Billion Corpus [2] for English and Russian News Crawl [3] for Russian.

### 4.2 Experimental Setting

We used the transformer model with copy-augmented architecture (Zhao et al., 2019) as the GEC model with almost the same hyperparameters. In particular, we set max-epoch $= 3$ for pre-training, and 15 for training. As an evaluation metric, we computed the precision, recall, and $\text{F}_{0.5}$ score for the CoNLL-2014 dataset and RULEC-GEC test set. Furthermore, we used the CoNLL-2013 (Ng et al., 2013) data and the RULEC-GEC dev data for development.

[1] `https://github.com/gutfeeling/word_forms`

[2] `https://www.statmt.org/lm-benchmark/`

[3] `http://www.statmt.org/wmt18/translation-task.html`

|  |  | CoNLL-2014 (En) |  |  | RULEC-GEC test (Ru) |  |  |
|---|---|---|---|---|---|---|---|
| System | Pseudo data | Prec. | Rec. | $F_{0.5}$ | Prec. | Rec. | $F_{0.5}$ |
| Random errors w/o Data selection (baseline) | 10M | 67.5 | 34.1 | 56.5 | 22.7 | 3.6 | 11.1 |
| Random errors w/ Data selection | 2M | 67.9 | 31.1 | 54.9 | 18.7 | 0.11 | 4.5 |
|  | 4M | 68.0 | 32.5 | 55.8 | 19.2 | 1.53 | 5.8 |
|  | 6M | 67.4 | 33.7 | 56.2 | 20.5 | 2.42 | 8.2 |
|  | 8M | 68.9 | 34.3 | 57.3 | 25.3 | 3.35 | 11.0 |
|  | 10M | 68.2 | **34.9** | 57.3 | 27.7 | 3.77 | 12.2 |
| Error type w/o Data selection | 10M | 69.2 | 34.2 | 57.5 | 41.1 | 12.4 | 28.1 |
| Error type w/ Data selection (proposed) | 2M | 67.5 | 31.3 | 54.8 | 32.8 | 2.5 | 9.7 |
|  | 4M | 68.8 | 33.1 | 56.6 | 37.2 | 6.7 | 19.5 |
|  | 6M | **70.0** | 33.5 | 57.5 | 44.2 | 11.9 | 28.6 |
|  | 8M | 68.5 | 34.6 | 57.2 | **49.0** | 15.0 | 33.7 |
|  | 10M | 69.1 | 34.5 | **57.6** | 48.6 | **16.8** | **35.2** |

Table 3: Results comparison of for each evaluated method. Best score in each column is indicated in bold.

As explained in Section 3.1, we trained the target LM to extract sentences from monolingual data using a part of the target side of the parallel data, where its domain matched the development data. We extracted the highest-scoring 10M sentences from the original monolingual datasets, One Billion Corpus, and Russian News Crawl, which have 30M and 80M sentences, respectively.

Furthermore, as discussed in Section 3.2, we generated pseudo data by incorporating pseudo errors into the monolingual corpus of each language. For noun case errors in Russian, we used a dictionary [4] containing noun inflections. We verified that the total number of pseudo errors in each experiment was similar to ensure a fair comparison. In our experiments, we compared the following three baselines to study the effects of pseudo errors and data selection in the monolingual corpus.

**Random errors w/o Data selection** In this approach, pseudo errors are added into randomly selected 10M monolingual data. The added errors include deleting, adding, and replacing randomly selected words, and shuffling the words in a sentence. This method corresponds to that of Zhao et al. (2019).

**Random errors w/ Data selection** First, we selected the top 10M sentences from the monolingual corpus using the LM scoring method described in Section 3.1. In our experiments, the amount of data is up to 10M sentences, increased by 2M sentences. In this approach, the process of adding pseudo errors is the same as in the Random errors w/o Data selection approach.

**Error type w/o Data selection** In this approach, we introduced pseudo errors to randomly selected 10M monolingual data, as described in Section 3.2.

**Error type w/ Data selection** This method is our proposed approach, where we combine the data selection and error type approaches.

### 4.3 Result

Table 3 lists the results for each system.

**Data selection.** When comparing the results obtained using the Random errors, we can evaluate the effect of the data selection method. For English, the random methods, which incorporated the data selection approach, perform better than the random method without it (56.5 → 57.3). In contrast, for Russian, similar improvements were noted for both approaches (11.1 → 12.2).

Furthermore, when comparing the results obtained using the error type, we confirmed that the data selection approach significantly improved GEC performance for Russian data. However, for the English data, no significant improvements for GEC performance were observed. Moreover, for the Russian data, we found that both precision and recall improved when using the error type-based approach (Precision: 41.1 → 48.6, Recall: 12.4 → 16.8).

**Error types.** When comparing random and error type w/ data selection approaches, we observed the effect of pseudo data containing pseudo errors based on learners' error types in GEC performance. For the English data, the improvement is

[4] http://opencorpora.org/?page=downloads

| System | Sentence |
|---|---|
| Source Sentence | We know each others' status, changements and so on through the social media. |
| Gold Sentence | We know each others' status, changes and so on through the social media. |
| Random w/ Data selection | We know each others' status, changements and so on through the social media. |
| Error type w/ Data selection | We know each others' status, changes and so on through the social media. |
| Source Sentence | Besides, we can make more friends by such interactions when our friends ... |
| Gold Sentence | Besides, we can make more friends through such interactions when our friends ... |
| Random w/ Data selection | Besides, we can make more friends through such interactions when our friends ... |
| Error type w/ Data selection | Besides, we can make more friends with such interactions when our friends ... |
| Source Sentence | В сочинение было много ошибок. |
| Gold Sentence | В сочинении было много ошибок. (En: There were many mistakes in the essay.) |
| Random w/ Data selection | В сочинение было много ошибок. |
| Error type w/ Data selection | В сочинении было много ошибок. |

Table 4: Comparison of system outputs in English and Russian. Examples on the top indicate those word form errors that were successfully corrected, while those on the middle indicate preposition errors that were not successfully corrected. Those on the bottom indicate noun case errors that were successfully corrected in Russian.

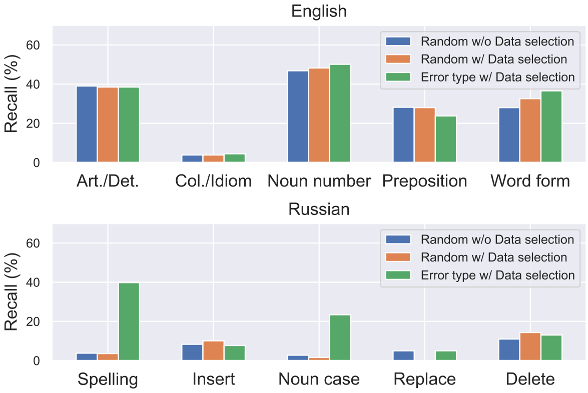

Figure 2: Comparison of recall for each error type. All systems were input with 10M pseudo data sentences.

not large. In contrast, for Russian data, the proposed method achieved the same level of accuracy using only one-third of the parallel corpus (8.23 → 9.68). Moreover, using the same amount of data, the score was almost tripled (12.2 → 35.2).

## 5 Analysis

**Error type.** Figure 2 shows the recall for each error type. We selected error types that most commonly appear in the development data.

For English data, the recall was comparable for all error types. Regarding error types other than preposition errors, an equal or improved recall was realized. In contrast, for preposition errors, the recall reduced significantly. It seems that this degradation in the recall can be attributed to the method used to add preposition errors in our study. In particular, we only considered replacement for preposition error generation, and not deletion or insertion. We believe this problem could be handled by generating preposition errors via insertion and deletion as well.

For Russian data, recall improved significantly for spelling and noun error cases. Note that these two error types are not considered explicitly during random error generation. In contrast, recalls for other error types are approximately comparable because the errors were generated using the same approach. Therefore, overall, we observed that the approach significantly improved by considering error types that could not be obtained using random error generation.

**Example.** Table 4 lists the output examples of two systems: Random errors w/ data selection and error type w/ data selection. Words in red indicate errors in the sentence, while those in blue indicate correct words.

At the top of Table 4, we present an instance of a word form error that was corrected using the proposed method. In particular, the random method outputted the input sentence as it stands. Conversely, the proposed method corrected the word form error by considering other word forms.

Furthermore, in the middle of Table 4, we present an output example wherein preposition errors were left uncorrected by the proposed method. In particular, the random method corrected the preposition error in an appropriately; however, our proposed method failed in performing the task. This difference in results is due to the limitations we posed on the dataset for the replacement to generate realistic pseudo errors. Thus, this example suggests that the recall degradation for preposition errors was caused by restricting the confusion set too strictly.

Finally, in the bottom of Table 4, we present an instance of a noun case error in Russian. The word “сочинение” is a neuter noun, and this case inflection of the word represents nominative or accusative case. When this word is used with the preposition “В”, meaning English “in” in this example, it is necessary to change the case to prepositional case (сочинение → сочинении). From this example, our proposed method can correct noun case error, while the random method cannot correct them.

As an overall tendency of Russian noun case errors, the random method often outputted the input sentence as it is, according to our observation of the outputs, or it outputted a completely different word.

As a case of failure to correct, in our proposed method, we confirmed a tendency that the method changed case inflections to the wrong ones.

## 6 Conclusions

In this study, we studied the effect of pseudo data obtained using two approaches. In particular, we confirmed that combining data selection and realistic error injection approaches to obtain pseudo data improved the $F_{0.5}$ scores. Moreover, we analyzed the recall for each error type. Based on our experimental results, we observed that the recall for error types considered in our study improved or were comparable.

## Acknowledgements

We would like to thank Yangyang Xi for granting permission to use text from Lang-8. We also thank Aizhan Imankulova and the anonymous reviewers for their insightful comments. This work was supported by JSPS KAKENHI, Grant Numbers JP19K12099 and JP19KK0286.

## References

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In *BEA*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *BEA*.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2019. Minimally-augmented grammatical error correction. In *W-NUT*.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *BEA*.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. In *EMNLP*.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *EMNLP*.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *COLING*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *ACL*.

Jakub Náplava and Milan Straka. 2019. Grammatical error correction in low-resource scenarios. In *W-NUT*.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *CoNLL*.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *TACL*.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *ACL*.