WikiComments: Leveraging Revision Comments to Extract Annotated Grammatical Correction Data from Wikipedia

Anonymous ACL submission

Abstract

We present *WikiComments*[1] a data extraction method which leverages the revision comments of Wikipedia edits to extract grammatical error correction training data. *WikiComments* improves the previous Wikipedia extraction method by only extracting data which are explicitly grammatical in nature. Our method produces larger quantities of data—up to 143% more—than existing benchmarks in languages such as German and Russian. We show that augmenting Korean training data with our extracted data leads to state-of-the-art results. Additionally, we show that augmenting minimal amounts of gold annotated data with *WikiComments* improves performance on up to 92% of German error types.

1 Introduction

Grammatical Error Correction (GEC) is currently dominated by English models and datasets. Research on languages other than English is considered a low-resource task given the current scarcity of evaluation benchmarks and annotated training data (Bryant et al., 2022). For instance, large public training datasets for English GEC such as Lang-8 (Mizumoto et al., 2011) contain more than a million samples, while for German there is currently only one dataset Falko-MERLIN (Boyd, 2018) with 19K training samples. Annotating more training data for these languages is an expensive process due to the need to find experts in each language, which hinders non-English GEC research.

To close this data scarcity gap, existing literature focuses on either *generating* artificial grammatical errors (Náplava and Straka, 2019) or *extracting* already available grammatical correction data from online sources (Boyd, 2018; Grundkiewicz and Junczys-Dowmunt, 2014) or a combination of both approaches (Lichtarge et al., 2019; Grundkiewicz et al., 2019). These approaches come with their respective advantages and drawbacks. The main approaches for generating synthetic data corrupt error-free sentences by introducing grammatical errors using either rule-based corruption (Náplava and Straka, 2019) or machine translation (Lichtarge et al., 2019), Large Language models have also been used for synthetic dataset generation for other tasks (Gupta et al., 2023). However, artificial errors from these methods do not necessarily reflect the errors humans make in the context of the original text. They also rely on either a set of grammatical corruption rules or access to a large language model that supports the specific language—which can be expensive to create or run, respectively.

An alternative approach involves extracting pairs of ungrammatical and their respective corrections from public edit logs. The primary source of extracted GEC data is Wikipedia, given its size, editorial quality, availability, and permissive licence (Boyd, 2018; Grundkiewicz and Junczys-Dowmunt, 2014; Lichtarge et al., 2019; Grundkiewicz et al., 2019). The main drawback is that these extraction approaches necessarily produce smaller amounts of training data (upper-bounded by the size of the available revision logs), compared to the virtually infinite space of synthetic data. However, the edits extracted can provide more realistic training data, since they reflect actual grammatical errors made by humans. The approach is also less costly than synthetic approaches since it does not involve text generation, only relatively simple processing of revision logs. For these reasons, we explore an extraction-based approach for building GEC data in this paper.

Given that revisions can be made for various reasons (e.g., correcting a factual inaccuracy), the current state-of-the-art extractive approaches for building GEC training corpora rely on various signals to determine whether a change is grammatical in nature. In this work, we make a simple—

[1] https://anonymous.4open.science/r/wikiComments-6C2A/

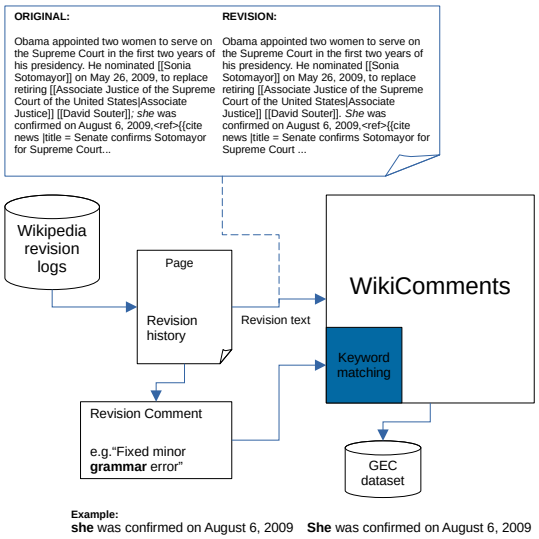

Figure 1: The WikiComments method extracts GEC training data by filtering the Wikipedia revision comments based on grammar-based keywords.

but important—observation: the comments associated with grammatical errors often indicate whether the change is due to a grammatical problem. Figure 1 provides an example. Building off *WikiEdits* (Grundkiewicz and Junczys-Dowmunt, 2014), we explore whether the Wikipedia edit *comments* provide a valuable signal for constructing training data for low-resource GEC through a keyword-based filtering approach that we call *WikiComments*.

We investigate if *WikiComments* yields consistently better grammatical error correction training data than the previous approach and in addition measure what is the quantity of filtered grammatical data we can extract compared to current datasets. We observe our filtering consistently helps the models and yields better performance across 3 different languages and 5 different datasets. This indicates that the revision comments contain helpful information for determining whether a revision is grammatical. Additionally, *WikiComments* can generate much larger quantities –up to 143% more– of data than the current gold benchmarks for languages with large numbers of edits in Wikipedia such as German and Russian.[2]

Following we fine-tune an mBART model with the *WikiComments* data since it was shown it is an appropriate baseline for GEC by Katsumata and Komachi (2020). We measure if we can improve performance compared to the gold datasets. We find that even though our extracted data are not a substitute for the gold data, we achieve SOTA results on the Kor-Union and Kor-Learner datasets by Yoon et al. (2023) when we augmented their training sentences with ours *WikiComments* data.

Finally, we investigate the performance when fine-tuning using a combination of WikiComments and low-resource gold sentences to determine if we can complement the dataset creation process by reducing the number of annotated training sentences needed for a robust GEC model. We achieved better performance in low-resource German, Russian and Korean instances when we augmented the training dataset with our data. Further error type analysis on German confirms that with our data we can achieve better performance on several error types the low resource data was unable to achieve on its own.

In summary, we introduce *WikiComments* a GEC data extraction method which leverages revision comments to detect grammatical error corrections. Our approach improves previous Wikipedia-based methods, can augment existing gold datasets and produce state-of-the-art results. We generate larger quantities than existing non-English benchmarks and we demonstrate the benefits of augmenting low-resource training datasets with our data.

2 Related Work

This section provides an overview of relevant literature. We explore literature for extracting existing grammatical correction data and generating artificial grammatical errors as well as literature combining both two methods.

2.1 Extracting Grammatical Data

Grundkiewicz and Junczys-Dowmunt (2014) introduced a novel grammatical error correction dataset called WikEd Error Corpus. Their corpus consists of approx. 12 million sentences extracted and filtered from Wikipedia edit history dumps. The authors iterate over two adjacent revisions of every page in the dump to construct the dataset. They removed unwanted edits such as cases of vandalism and markup and split the articles into sentences which made up their final English dataset.

Boyd (2018) developed a grammatical error correction system for German by combining a low-resource German gold corpus with sentence pairs extracted from Wikipedia using the tools developed by Grundkiewicz mentioned previously. The

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias

authors extend ERRANT (Bryant et al., 2017) to work with German data by simplifying the error types used. After training the multilayer convolutional GEC model proposed by Chollampatt and Ng (2018) on a combination of the gold German dataset (Falko-MERLIN) and German Wikipedia sentences they observe that augmenting a low resource dataset with Wikipedia-based sentences can improve model performance.

While both approaches leverage Wikipedia data, none of their filtering is satisfactory and their dataset contains a lot of edits that are not grammatical. Grundkiewicz and Junczys-Dowmunt (2014) while attempting to reduce noise do not ensure the extracted sentence pairs are of grammatical content. In contrast, Boyd (2018) filters edits by attempting to mirror the error distribution of the Falko-MERLIN dataset using ERRANT. While better than no filtering at all, it risks exposing evaluation signals to the training process. In addition, it requires a method to annotate the GEC errors and an existing benchmark for the filtering, components that are of limited availability depending on the language. We complement both of these approaches by filtering Wikipedia explicitly for GEC training data and not requiring the existence of a gold dataset or an annotation tool.

2.2 Generating Artificial Data

Grundkiewicz et al. (2019) propose an unsupervised method to generate artificial errors. They propose generating confusion sets based on the information from the vocabulary of an Aspell spellchecker. Their method uses those confusion sets to introduce errors into an error-free text by iterating through random words/characters and either substituting it with a random word/character from its confusion set, deleting the word/character or swapping it with an adjacent one.

Náplava and Straka (2019) present a novel grammatical error correction dataset for Czech (AKCES-GEC). In addition, they experiment with combining gold data with synthetic data to train a Transformer (Vaswani et al., 2017) model on multilingual GEC. Their synthetic data is generated by rule-based corruption of sentences from the WMT News Crawl (Bojar et al., 2017). Their results show that their method performs better than existing GEC systems.

Rothe et al. (2021) proposed leveraging large-scale language models to simplify the current complexity of synthetic approaches. They pre-train an XLL T5 (Raffel et al., 2020) model (gT5) on synthetically generated sentences in 101 languages and fine-tune the model on gold datasets for English, Russian, Czech and German. Then they leverage the model to clean the Lang-8 dataset and distil the model's knowledge on smaller models. This cleaned dataset cLang-8 contains processed sentences for English, German and Russian and they show that models fine-tuned on it can exhibit state-of-the-art performance. They argue this approach can simplify the training for GEC.

While these approaches exhibit competitive performance, their methods for generating artificial errors have drawbacks. Relying on a set of grammatical corruption rules results in corruption that may not accurately represent actual human errors. Additionally, they also depend on dictionaries with comprehensive confusion sets which might not exist for low-resource languages. Finally, obtaining sufficient computational resources for pre-training XXL models on hundreds of languages to distil GEC knowledge on more datasets –the cLang-8 dataset only contains training data for English, German, and Russian– makes these approaches either too costly for low-resource languages.

2.3 Combinations of Artificial and Extracted Data

Flachs et al. (2021) demonstrate that Lang8 and Wikipedia-based data whilst noisy can be beneficial to grammatical error correction research and that using even very small amounts of gold data for fine-tuning can yield good performance.

Finally, Lichtarge et al. (2019) explored two approaches to generate GEC datasets. The first involves extracting sentences from Wikipedia with minimal filtering, while the second introduces errors in Wikipedia sentences through round-trip translation. They demonstrate that GEC methods trained using either approach perform similarly.

While our focus is solely on extracting existing grammatical data rather than generating it artificially, existing literature shows that a combination of both approaches yields favourable results. Both papers also confirm the validity of Wikipedia as a reliable GEC data source. This paper aims to enhance the existing Wikipedia extraction approach, potentially contributing to further improvements to methods leveraging combinations of extraction and generation of GEC data.

2.4 Other Approaches

Katsumata and Komachi (2020) proposed using

monolingual (Lewis et al., 2020) and multilingual BART (Liu et al., 2020) model as a pretrained GEC model that can be easily used as a baseline. Their results show that an mBART model fine-tuned in German and Czech can achieve competitive results using only gold data and can be used as a baseline for several languages. We build on their work by using mBART for our *WikiComments* experiments.

Fang et al. (2023) proposed using chain-of-thought and ChatGPT as a GEC method. They evaluate 3 different languages (English, German and Chinese) and demonstrate that ChatGPT can perform well in low-resource languages. One limitation of Fang et al. (2023) is its approach is entirely reliant on the external API of OpenAI whereas we aim to provide a GEC solution that can be utilized even in low resource settings.

3 Methodology

The *WikiComments* sentence extraction method parses through dumps of Wikipedia revision history based on the WikiEdits method of Grundkiewicz et al (Grundkiewicz and Junczys-Dowmunt, 2014) to extract grammatical error correction training data as shown in Figure 2. *WikiComments* extracts parallel sentences by filtering Wikipedia revision history dumps. To reduce noise, the revisions are filtered to remove reverted revisions, markup, code and any revisions where the only change is numerical. Unlike the *WikiEdits* approach, *WikiComments* leverages the revision comments to heuristically exclude edits where the comment does not indicate any change to correct grammatical errors i.e. contain any word from a list of keywords indicating grammatical fixes such as “grammar" or “typo". For a full list of the keywords check Appendix A. The resulting sentence pairs can be used to fine-tune neural models on a GEC task. For a list of sample sentence pairs filtered out by *WikiComments* compared to *WikiEdits* check Appendix B.

4 Experimental Setup

This section details our experiment setup used to answer the following research questions:

- **RQ1**: How does the quantity of the data extracted with *WikiComments* compare to the gold datasets and the *WikiEdits* approach?
- **RQ2**: Does filtering with *WikiComments* improve the quality of the training data extracted from Wikipedia?

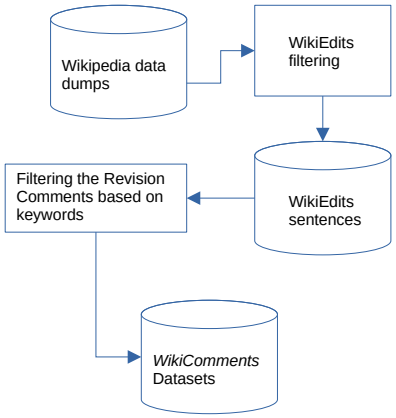

Figure 2: Filtering process to extract GEC data from Wikipedia

- **RQ3**: Does fine-tuning on the *WikiComments* data improve performance compared other non-Wikipedia based approaches?
- **RQ4**: Does augmenting low resource gold data with *WikiComments* data improve model performance?

4.1 Models

Motivated by the performance of mBART in Katsumata and Komachi (2020) we experiment with the 680M-parameter mBART developed by Liu et al. (2020). mBART is a sequence-to-sequence auto-encoder that was shown to can perform well in both supervised and unsupervised machine translation. We fine-tune mBART using the fairseq toolkit by Ott et al. (2019).

4.2 Languages

Since there is substantial human-annotated data for languages such as English, we focus on three low-resource languages with comparably smaller annotated benchmarks. Our selection includes languages from different language families to measure how well our *WikiComments* extraction process works. The criteria for our selection were the availability of existing annotated benchmarks for evaluation, the presence of a dedicated Wikipedia, and the availability of a spaCy pipeline for pre-processing. Therefore, we focus on German (Indo-European), Korean (Koreanic) and Russian (Indo-European).

4.3 Datasets

We use the Falko-MERLIN Dataset by Boyd (2018) for our experiments on German. For our experiments on Russian, we use the RULEC-GEC dataset

by Rozovskaya and Roth (2019). Finally, for Korean, we use the Kor-Native, Kor-Learner and Kor-Union datasets by Yoon et al. (2023).

4.4 Evaluation

We evaluate our models using a modified version of the *MaxMatch* (M^2) method by Dahlmeier and Ng (2012) that we modified to use spaCy tokenization to be consistent with the sentence segmentation of our sentence extraction. Finally, we performed an error type analysis on German to identify which error types' performance was improved when we combined the low resource data with our data. We use the ERRANT modified by Boyd (2018) to get the performance of each error type. All our reported M^2 results are from single runs.

4.5 Filtering the dumps from Wikipedia

To answer RQ1 and RQ2 we first extract datasets of sentence pairs from the Wikipedia data dumps. For each of the three languages, we produce two Wikipedia-based datasets, a *WikiComments* (WC) and a *WikiEdits* (WE) one based on the process explained in Section 3. The WikiComments datasets contain the sentence pairs extracted with our methodology leveraging the **revision comment** filter to obtain annotated grammatical data and the WikiEdits datasets contain the sentence pairs extracted without the comment filter. Our filter tries to match keywords that relate to changes in grammatical content. We generated these keywords by manual inspection of their respective Wikipedia revision history and translating common English terms such as "grammar", "grammatical" etc to the target language. We sample 10K, 20K, 50K, and 100K sentences for German and Russian and only 10K and 20K samples for Korean due to the small number of sentences extracted from Wikipedia. We fine-tuned an mBART model by iterating on an increasing amount of training sentences to measure whether the comment-based filtering yields better GEC training data and how many sentence pairs we need for optimal performance.

4.6 Complementing Low Resource Data

To answer RQ3 and RQ4 we fine-tune mBART solely on our *WikiComments* datasets and on different combinations of the optimal amount of *WikiComments* grammatical sentences found in the previous experiments and the existing gold datasets. We run experiments combining the entire gold dataset plus training sentences extracted using our WikiComments approach to investigate if we can augment the entire gold collection. We additionally try sampling training sentences from the gold datasets to identify if our Wikipedia sentences can complement low-resource scenarios by augmenting the samples with our datasets. We sample 1K, 2.5K, 5K and 10K gold sentence pairs for Kor-Union and Falko-MERLIN and 1K and 2.5K (due to the small size of the training split) from RULEC and fine-tune mBART on combinations of the gold samples and the WikiComments data.

Language / Dataset	Gold	WC	WE
Korean	156K	24K	2.8M
German	24K	1.2M	33M
Russian	12K	1.8M	19M

Table 1: Number of total sentences extracted using WikiComments (WC) compared to total sentences of WikiEdits (WE) and the gold datasets we used for evaluation.

5 Results and Discussion

This section details our experiments to answer our research questions. Each subsection contains our findings and discussion for each research question.

5.1 Comparing quantity of *WikiComments* data with other approaches

Table 1 shows the number of sentences extracted with WikiEdits and WikiComments and the total sentences of the gold datasets we used for evaluation. Our approach generates less data compared to *WikiEdits* but as shown in the previous section we produce data of better quality. We can verify this by the small portion of grammatical errors in the sentences extracted using *WikiEdits*. Only 4% of German, 9% of Russian and less than 1% of the Korean sentences extracted using *WikiEdits* are kept by *WikiComments* which indicates the small number of grammatical errors in their datasets. Compared to the gold datasets we produce many more – up to 143 % – sentences in German and Russian than the two existing gold datasets, Falko-Merlin and RULEC respectively. Additionally, our Korean sentences are fewer in number compared to the sentences in the Kor-Union dataset.

The number of sentences extracted with *WikiComments* coincides with the relative ranking of each language based on the number of edits on Wikipedia, with German and Russian at 2nd and

Figure 3: Comparison of German *WikiComments* and *WikiEdits* mBART runs evaluated on Falko-MERLIN

Figure 4: Comparison of Russian *WikiComments* and *WikiEdits* mBART runs evaluated on RULEC

6th rank and Korean at 19th. The fewer edits in total exist in Wikipedia the smaller the pool of edits we can filter to extract the *grammatical* edits from. Conversely, the difference between our Korean WikiComments and Kor-Union can be explained by the fact that Kor-Union is made up of three datasets: Kor-Native, Kor-Learner and Kor-Lang8.

In conclusion, the *WikiEdits* approach generates larger datasets, but most of their content lacks grammatical error corrections. In contrast, our approach contains only revisions addressing grammatical errors and produces larger quantities of data compared to most gold datasets.

5.2 Measuring the impact of the comment filtering

We can measure the data quality by testing the effectiveness of mBART models trained on compared with the effectiveness when trained on the WikiEdits datasets. Figure 3 shows the F0.5 performance of mBART models trained on *WikiComments* and *WikiEdits* on a variety of training sample sizes. We can see a clear performance improvement with our *WikiComments* extracted data on German. Even on the smallest training sample of 10K sentence pairs the mBART mode fine-tuned on *WikiComments* outperforms all *WikiEdits* experiments.

We observe similar findings for the Korean *WikiComments* datasets, when evaluated on Kor-Union, Kor-Learner & Kor-Native, which can also be seen in Figures 5 & 6 with the difference being more notable on the Kor-Learner dataset. From Figure 4 we can also similar trends as the other languages with the *WikiComment* runs performing better than the *WikiEdits* runs.

Overall, we can answer **RQ2** that leveraging the revision comments for our *WikiComments* data extraction approach improves the existing *WikiEdits* approach and can produce better quality grammatical error correction training data.

5.3 Comparing filtered Wikipedia data with Gold Datasets

Table 2 presents our findings when we fine-tuned mBART with our *WikiComments* datasets. For German and Russian our *WikiComments* datasets perform worse on their own and degrade the performance of mBART when combined with the gold data compared to just using the gold datasets. This can indicate that while there's a notable improvement over *WikiEdits*, *WikiComment* data are not a substitute for expertly annotated datasets on these languages. On the other hand, when we augmented the German gold data with ours we outperformed the ChatGPT approach by Fang et al. (2023) a method which was fine-tuned on significantly larger amounts of data compared to ours.

We see similar findings for Korean, where the *WikiComments* datasets are not a substitute for the respective gold training data. One reason for this is that the model used by Yoon et al. (Yoon et al., 2023) KoBART was pre-trained only on Korean data whereas mBART was pretrained on cc25(Wenzek et al., 2020; Conneau et al., 2020) a corpus of 25 languages from different languages families which might have a different impact on model performance compare to a model exposed only on a single language.

However, unlike our experiments on German and Russian, combining the gold training with ours

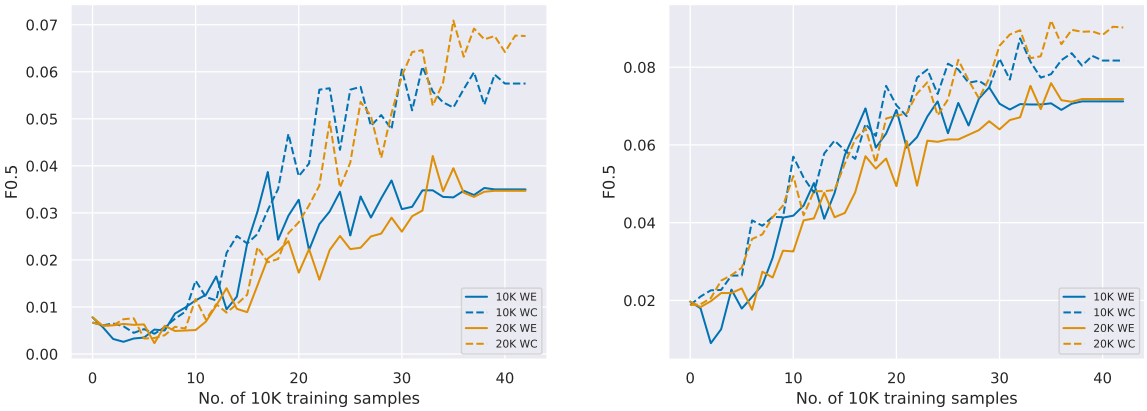

Figure 5: Comparison of Korean *WikiComments* and *WikiEdits* mBART runs evaluated on Kor-Learner and Kor-Union

Model	FT Data	F0.5
	German	
Transformer	FM + WMT	**0.737**
mT5 Large	cLang-8	0.701
mBART	FM	0.690
mBART (ours)	FM + 50K WC	*0.668*
ChatGPT	Various	0.635
MLConv	FM + 1M WE	0.452
mBART (ours)	50K WC	*0.384*
MLConv	1M WE	0.158
	Russian	
Transformer	RULEC + WMT	**0.502**
mT5 Large	cLang-8	0.276
mBART	RULEC	0.154
mBART (ours)	50K WC	*0.117*
mBART (ours)	RULEC + 50K WC	*0.125*
	Korean	
	Kor-Union	
mBART (ours)	KU + 19K WC	***0.333***
KoBART	KU	0.317
mBART (ours)	19K WC	*0.089*
	Kor-Learner	
mBART (ours)	KL + 19K WC	***0.453***
KoBART	KL + KU	0.410
KoBART	KL	0.376
mBART (ours)	19K WC	*0.068*
	Kor-Native	
KoBART	KN + KU	**0.736**
KoBART	KN	0.705
mBART (ours)	KN + 19K WC	*0.551*
mBART (ours)	19K WC	*0.210*

Table 2: Comparison of mBART trained on WikiComments (WC) with approaches with different fine-tuning data. Results are grouped by evaluation dataset: Falko-MERLIN, RULEC and Kor-Union, Kor-Learner and Kor-Native. The best-performing approach is shown in **bold** and our results are shown in *italics*.

Figure 6: Comparison of Korean *WikiComments* and *WikiEdits* mBART runs evaluated on Kor-Native

yields **state-of-the-art** performance on both Kor-Union and Kor-Learner. In conclusion, while our *WikiComments* datasets are not a substitute for the current gold datasets, our results on Kor-Union and Kor-Learner demonstrated that by combining the gold training and the *WikiComments* data we can obtain SOTA results. This is an indication that our dataset contains different types of grammatical errors than the gold training sentences and combining them exposes mBART to a larger pool of errors and allows us to obtain better GEC performance.

To answer **RQ3**, fine-tuning solely on *WikiComments* data does not improve the performance compared to the current gold datasets. However, the results on the two Korean datasets when augmented by the *WikiComments* show that there is utility to our method. Investigation on the optimal utilisation of *WikiComments* datasets is left for future work.

5.4 Complementing low resource gold datasets with Wikipedia data

Sample Size	Gold Only	+ WC	δ
German			
10K	0.618	0.602	-0.016
5K	0.576	0.571	-0.005
2.5K	0.496	0.519	**+0.023**
1K	0.258	0.478	**+0.220**
Russian			
2.5K	0.064	0.031	-0.033
1K	0.010	0.031	**+0.021**
Korean			
Kor-Union			
10K	0.243	0.223	-0.020
5K	0.124	0.083	-0.041
2.5K	0.125	0.079	-0.046
1K	0.035	0.162	**+0.127**

Table 3: Comparison of F0.5 performance of mBART fine-tuned only on the gold sample and mBART fine-tuned on the gold sample and WikiComments

Table 3 presents the findings of our investigation on whether our data can help research on low-resource languages. We draw samples of decreasing size from the gold datasets and compare the sample's performance with an mBART instance fine-tuned on a combination of the gold sample and data extracted by *WikiComments*. Across all three languages, our dataset augmentations consistently perform better than the 1K samples on their own. Notably, we observe an improvement in performance on the 2.5K German sample. An interesting observation is the massive decline in performance in all Russian samples compared to the performance on the entire RULEC dataset, as reported in Section 5.3, something not observed in the samples of Kor-Union or Falko-MERLIN.

Table 4 shows the results of our German error type analysis. These results validate our earlier findings on German. Specifically, 92% of error types show improvement for the 1K sample, 64% for the 2.5K sample, and 48% for the 5K sample. Notably, seven error types —AUX:FORM, CONJ, NOUN, OTHER, PNOUN, SCONJ, and SPELL— consistently exhibit improvements across all three sample sizes. This shows the benefits of augmenting a small gold dataset with our *WikiComments* dataset, especially for languages where annotating gold sentences is resource-intensive. To answer **RQ4**, we find that augmenting low-resource gold data benefits small collections of 1K sentence pairs.

Error Type	1K	2.5K	5K
ADJ	**+0.080**	-0.018	**+0.050**
ADJ:FORM	**+0.184**	**+0.031**	-0.042
ADP	**+0.068**	-0.068	**+0.011**
ADV	**+0.042**	**+0.010**	-0.062
ADV:FORM	0	-0.769	-0.049
AUX	**+0.134**	-0.006	-0.007
AUX:FORM	**+0.134**	**+0.030**	**+0.020**
CONJ	**+0.306**	**+0.256**	**+0.055**
DET	**+0.236**	**+0.107**	-0.026
DET:FORM	**+0.096**	-0.038	**+0.030**
MORPH	**+0.220**	**+0.088**	-0.026
NOUN	**+0.191**	**+0.026**	**+0.096**
NOUN:FORM	**+0.063**	-0.028	-0.066
ORTH	**+0.042**	**+0.008**	-0.056
OTHER	**+0.213**	**+0.041**	**+0.037**
PART	**+0.198**	**+0.088**	-0.029
PNOUN	**+0.159**	**+0.149**	**+0.230**
PRON	**+0.198**	**+0.031**	-0.053
PRON:FORM	**+0.193**	-0.037	-0.055
PUNCT	-0.006	**+0.039**	**+0.035**
SCONJ	**+0.123**	**+0.038**	**+0.002**
SPELL	**+0.251**	**+0.077**	**+0.038**
VERB	**+0.097**	-0.026	**+0.052**
VERB:FORM	**+0.071**	-0.032	-0.075
WO	**+0.119**	**+0.004**	-0.044

Table 4: Delta of error type F0.5 of Falko-MERLIN test evaluation on mBART sample vs sample + WC

Further error analysis on German confirms the effectiveness of our augmentation, resulting in improved performance when evaluated on both the entire test collection and individual error types.

6 Conclusion

This paper introduces *WikiComments*, a data extraction method that leverages the revision comment history of Wikipedia to obtain "silver" grammatical error correction data. The approach filters out Wikipedia revisions without comments mentioning edits of grammar errors. Our experiments demonstrate the consistently better performance of *WikiComments* compared to the previous *WikiEdits* method across German, Korean, and Russian datasets of various sizes. While not a substitute for "gold" annotated data, *WikiComments* can be used to augment them, achieving state-of-the-art results on Kor-Learner and Kor-Union. Additionally, *WikiComments* proves helpful for augmenting low-resource training. In summary, *WikiComments* provides a reliable and efficient means of producing GEC data for low-resource tasks by leveraging publicly available edits.

7 Limitations

This section outlines the limitations of this paper. First, We only use a single neural model to evaluate our extraction method, mBART which is a potential limitation in investigating whether our findings generalise across neural models with different architectures. Another limitation is that we only evaluate our method on three languages, two sharing the same language family. Finally, in our experiments extracting parallel sentences from Wikipedia dumps takes a considerable amount of time even though we took sufficient steps to optimise the extraction process.

References

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 169–214. Association for Computational Linguistics.

Adriane Boyd. 2018. Using wikipedia edits in low resource grammatical error correction. In *Proceedings of the 4th Workshop on Noisy User-generated Text, NUT@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 79–84. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 793–805. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2022. Grammatical error correction: A survey of the state of the art. *CoRR*, abs/2211.05166.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5755–5762. AAAI Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 568–572. The Association for Computational Linguistics.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? A comprehensive evaluation. *CoRR*, abs/2304.01746.

Simon Flachs, Felix Stahlberg, and Shankar Kumar. 2021. Data strategies for low-resource grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EACL, Online, April 20, 2021*, pages 117–122. Association for Computational Linguistics.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings*, volume 8686 of *Lecture Notes in Computer Science*, pages 478–490. Springer.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2019, Florence, Italy, August 2, 2019*, pages 252–263. Association for Computational Linguistics.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. 2023. Targen: Targeted data generation with large language models. *CoRR*, abs/2310.17876.

Satoru Katsumata and Mamoru Komachi. 2020. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 827–832. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy,

Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3291–3301. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated japanese error correction of second language learners. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pages 147–155. The Association for Computer Linguistics.

Jakub Náplava and Milan Straka. 2019. Grammatical error correction in low-resource scenarios. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 346–356. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 702–707. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically-rich languages: The case of russian. *Trans. Assoc. Comput. Linguistics*, 7:1–17.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4003–4012. European Language Resources Association.

Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. Towards standardizing Korean grammatical error correction: Datasets and annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6742, Toronto, Canada. Association for Computational Linguistics.

A Revision Filter Keywords

A.1 German Keywords

grammatikkorr, grammatik, tippfehler, grammatikalisch, grammatikfehler, grammatika, tippfehlerkorrektur, tippfehlerkorrigieren, tippfehlerkorrigiert, tippfehlerkorrigierte

A.2 Korean Keywords

오식|철자가, 틀리다|맞춤법, 오류|문법|문법, 오류|맞춤법,

A.3 Russian Keywords

опечаткам, опечатками, опечатках, опечаткой, опечаткам, опечатками, опечатках, орфография, орфографический, орфографическая, орфографические, орфографических, орфографических, пунктуация, пунктуационный, пунктуационная, пунктуационные, пунктуационных, пунктуационных, грамматика, грамматический, опечатка, опечатки, опечаток, опечатке, опечаткой

B WikiComments Filtering Examples

Below are some samples of queries the WikiEdits extracted compared to the same sample as filtered with WikiComments.

B.1 German

B.1.1 WikiEdits

Alan Smithee ist eigentlich kein Regisseur, sondern ein Anagramm von "The Alias Man", was bedeutet, daß Filme, in denen Alan Smithee Regie geführt hat, dem eigentlichen Regisseur so peinlich waren, daß er seinen Namen nicht dafür aufs Spiel setzen wollte. Alan Smithee ist eigentlich kein Regisseur, sondern ein Anagramm von "The Alias Men", was bedeutet, daß Filme, in denen Alan Smithee Regie geführt hat, dem eigentlichen Regisseur so peinlich waren, daß er seinen Namen nicht dafür aufs Spiel setzen wollte.

Alan Smithee ist eigentlich kein Regisseur, sondern ein Anagramm von "The Alias Men", was bedeutet, daß Filme, in denen Alan Smithee Regie geführt hat, dem eigentlichen Regisseur so peinlich waren, daß er seinen Namen nicht dafür aufs Spiel setzen wollte. Alan Smithee ist eigentlich kein Regisseur, sondern ein Anagramm von "The Alias Men", was bedeutet, dass Filme, in denen Alan Smithee Regie geführt hat, dem eigentlichen Regisseur so peinlich waren, dass er seinen Namen nicht dafür aufs Spiel setzen wollte.

Wenn ein Film aber nachweislich stark gegen den Willen des Regisseurs verändert wurde, dann darf das Pseudonym Allen Smithee verwendet werden, und nur dieses. Wenn ein Film aber nachweislich stark gegen den Willen des Regisseurs verändert wurde, dann darf das Pseudonym Allan Smithee verwendet werden, und nur dieses.

1997 kam die Parodie (dt: Fahr' zur Hölle, Hollywood) in die Kinos, damit war das Pseudoym offensichtlich enttarnt und damit weniger nützlich. 1997 kam die Parodie (dt: "Fahr' zur Hölle, Hollywood") in die Kinos, damit war das Pseudoym offensichtlich enttarnt und damit weniger nützlich.

Wenn ein Film aber nachweislich stark gegen den Willen des Regisseurs verändert wurde, dann darf das Pseudonym Alan Smithee verwendet werden, und nur dieses. Wenn ein Film nachweislich stark gegen den Willen des Regisseurs verändert wurde, dann darf das Pseudonym Alan Smithee verwendet werden, und nur dieses.

B.1.2 WikiComments

1997 kam die Parodie (dt: "Fahr' zur Hölle, Hollywood") in die Kinos, damit war das Pseudoym offensichtlich enttarnt und damit weniger nützlich; 1997 kam die Parodie (dt: "Fahr' zur Hölle, Hollywood") in die Kinos, damit war das Pseudonym offensichtlich enttarnt und damit weniger nützlich;

B.2 Korean

B.2.1 WikiEdits

스위스의요한베르누이와프랑스의수학자들의활약이눈부시다. 스위스의베르누이일가와프랑스의수학자들의활약이눈부시다.

고대수학을크게발전시킨나라로는이집트,인도,그리스, 중국등이있다. 고대수학을크게발전시킨나라로는이집트, 인디아, 그리스, 중국등이있다.

후에미적분학을누가먼저창안하였는지에대한논쟁이있었으나현재는두사람이독립적으로그업적을이루었다는것이밝혀졌다. 후에미적분학을누가먼저창안하였는지에대한논쟁이있었으나현재는두사람이독립적으로그업적을이루었다는것이밝혀졌다.

그외오일러와더불어변분학을창시한라그랑주, 천체의운동을수학적으로규명한라플라스, 타원함수론의선구자였던르장드르, 화법기하학을창시한몽주등이있다. 그외오일러와더불어변분학을창시한라그랑주, 천체의운동을수학적으로규명한라플라스, 타원함수론의선구자였던르장드르, 화법기하학을창시한몽주등이있다.

방법서설을지은철학자데카르트는해석기하학의창시자로불후의이름을남기고있다. 《방법서설》을지은철학자데카르트는해석기하학의창시자로불후의이름을남기고있다.

B.2.2 WikiComments

고대수학을크게발전시킨나라로는이집트,인도,그리스, 중국등이있다. 고대수학을크게발전시킨나라로는이집트, 인디아, 그리스, 중국등이있다.

후에미적분학을누가먼저창안하였는지에대한논쟁이있었으나현재는두사람이독립적으로그업적을이루었다는것이밝혀졌다. 후에미적분학을누가먼저창안하였는지에대한논쟁이있었으나현재는두사람이독립적으로그업적을이루었다는것이밝혀졌다.

그외오일러와더불어변분학을창시한라그랑주, 천체의운동을수학적으로규명한라플라스, 타원함수론의선구자였던르장드르, 화법기하학을창시한몽주등이있다. 그외오일러와더불

어변분학을창시한라그랑주, 천체의운동을수학적으로규명한라플라스, 타원함수론의선구자였던르장드르, 화법기하학을창시한몽주등이있다.

B.3 Russian

B.3.1 WikiEdits

3 августа 1940 Верховный Совет СССР удовлетворил эту «просьбу». 3 августа 1940 Верховный Совет СССР удовлетворил эту просьбу.

Название «Литва» (Lituae) впервые упомянуто в Кведлинбургских летописях в 1009. Название «Литва» (Lituae) впервые упомянуто в Кведлинбургских летописях в 1009 году.

Свидетельством существования таких догосударственных объединений считается договор 1219 между галицко-волынскими князьями и 21 литовским князем. Свидетельством существования таких догосударственных объединений считается договор 1219 года между галицко-волынскими князьями и 21 литовским князем.

Вскоре они покорили Пруссию и Ливонию , а оставшиеся непокорёнными земли объединились под защитой Литвы. Вскоре они покорили Пруссию и Ливонию, а оставшиеся непокорёнными земли объединились под защитой Литвы.

Mindaugas, 1236—1263) принял католическое крещение в 1251 и был коронован 6 июля 1253. Mindaugas, 1236—1263) принял католическое крещение в 1251 году и был коронован 6 июля 1253 года.

B.3.2 WikiComments

азвание «Литва» (Lituae) впервые упомянуто в Кведлинбургских летописях в 1009. Название «Литва» (Lituae) впервые упомянуто в Кведлинбургских летописях в 1009 году.

Свидетельством существования таких догосударственных объединений считается договор 1219 между галицко-волынскими князьями и 21 литовским князем. Свидетельством существования таких догосударственных объединений считается договор 1219 года между галицко-волынскими князьями и 21 литовским князем.

Вскоре они покорили Пруссию и Ливонию , а оставшиеся непокорёнными земли объединились под защитой Литвы. Вскоре они покорили Пруссию и Ливонию, а оставшиеся непокорёнными земли объединились под защитой Литвы.

Mindaugas, 1236—1263) принял католическое крещение в 1251 и был коронован 6 июля 1253. Mindaugas, 1236—1263) принял католическое крещение в 1251 году и был коронован 6 июля 1253 года.

C ERRANT Full Analysis

This section shows the full ERRANT output for all the low resource German experiments in Section 5.4.

Error Type	P	R	F0.5
ADJ	0.160	0.087	0.137
ADJ:FORM	0.640	0.201	0.445
ADP	0.154	0.048	0.107
ADV	0.0	0.0	0.0
ADV:FORM	1.0	0.0	0.0
AUX	0.375	0.146	0.285
AUX:FORM	0.526	0.124	0.319
CONJ	0.23	0.071	0.027
DET	0.041	0.050	0.043
DET:FORM	0.659	0.422	0.591
MORPH	0.636	0.083	0.272
NOUN	0.033	0.014	0.026
NOUN:FORM	0.789	0.187	0.480
ORTH	0.453	0.137	0.311
OTHER	0.038	0.065	0.041
PART	0.625	0.091	0.287
PNOUN	0.0	0.0	0.0
PRON	0.014	0.039	0.016
PRON:FORM	0.267	0.156	0.233
PUNCT	0.673	0.402	0.593
SCONJ	0.065	0.036	0.056
SPELL	0.446	0.079	0.231
VERB	0.375	0.054	0.171
VERB:FORM	0.818	0.074	0.273
WO	0.472	0.126	0.305

Table 5: ERRANT error type analysis for mBART fine-tuned only on 1K sample of Falko-MERLIN

Category	P	R	F0.5
ADJ	0.2609	0.1304	0.2174
ADJ:FORM	0.7699	0.364	0.6295
ADP	0.5161	0.048	0.1751
ADV	0.1111	0.0119	0.0417
ADV:FORM	1.0	0.0	0.0
AUX	0.5316	0.227	0.4192
AUX:FORM	0.5714	0.2469	0.4525
CONJ	0.5	0.1429	0.3333
DET	0.4493	0.1107	0.2788
DET:FORM	0.8125	0.4259	0.6876
MORPH	0.7727	0.2012	0.4928
NOUN	0.2647	0.1268	0.2174
NOUN:FORM	0.7045	0.2831	0.5429
ORTH	0.4	0.2384	0.3522
OTHER	0.3319	0.1306	0.2537
PART	0.8333	0.1818	0.4854
PNOUN	0.1667	0.1333	0.1587
PRON	0.383	0.0779	0.2148
PRON:FORM	0.8462	0.1429	0.4264
PUNCT	0.6489	0.4255	0.5872
SCONJ	0.4286	0.0536	0.1786
SPELL	0.6442	0.241	0.4827
VERB	0.4848	0.0958	0.2676
VERB:FORM	0.7647	0.1074	0.3439
WO	0.625	0.1852	0.4237

Table 6: ERRANT error type analysis for mBART fine-tuned on 1K sample of Falko-MERLIN plus WikiComments

Category	P	R	F0.5
ADJ	0.2727	0.1304	0.2239
ADJ:FORM	0.748	0.3849	0.6293
ADP	0.4643	0.1171	0.2915
ADV	0.0769	0.0119	0.0368
ADV:FORM	0.8	0.6667	0.7692
AUX	0.5333	0.2595	0.4404
AUX:FORM	0.4722	0.2099	0.3778
CONJ	0.0667	0.2143	0.0773
DET	0.2126	0.1571	0.1986
DET:FORM	0.8057	0.6362	0.7649
MORPH	0.6364	0.2071	0.4499
NOUN	0.2333	0.0986	0.1832
NOUN:FORM	0.7788	0.3699	0.6378
ORTH	0.4557	0.301	0.4132
OTHER	0.2324	0.1478	0.2085
PART	0.5833	0.2545	0.4636
PNOUN	0.0	0.0	0.0
PRON	0.2333	0.0909	0.1777
PRON:FORM	0.7273	0.2078	0.4848
PUNCT	0.621	0.5271	0.5996
SCONJ	0.2727	0.0536	0.15
SPELL	0.6981	0.1775	0.44
VERB	0.6	0.1437	0.367
VERB:FORM	0.9412	0.1322	0.4233
WO	0.5667	0.2519	0.4533

Table 7: ERRANT error type analysis for mBART fine-tuned only on 2.5K sample of Falko-MERLIN

Category	P	R	F0.5
ADJ	0.24	0.1304	0.2055
ADJ:FORM	0.7554	0.4393	0.6604
ADP	0.55	0.0661	0.2231
ADV	0.1667	0.0119	0.0463
ADV:FORM	1.0	0.0	0.0
AUX	0.5281	0.2541	0.4344
AUX:FORM	0.5	0.2346	0.4077
CONJ	0.5	0.1429	0.3333
DET	0.4795	0.125	0.3059
DET:FORM	0.8132	0.5103	0.7269
MORPH	0.8261	0.2249	0.5382
NOUN	0.25	0.1268	0.2093
NOUN:FORM	0.7525	0.347	0.61
ORTH	0.4795	0.2828	0.4209
OTHER	0.3195	0.1323	0.249
PART	0.7778	0.2545	0.5512
PNOUN	0.1538	0.1333	0.1493
PRON	0.3393	0.0823	0.2088
PRON:FORM	0.7647	0.1688	0.4483
PUNCT	0.6943	0.4819	0.6381
SCONJ	0.5	0.0536	0.1875
SPELL	0.6591	0.2782	0.5174
VERB	0.5641	0.1317	0.3406
VERB:FORM	0.7083	0.1405	0.3917
WO	0.6078	0.2296	0.4572

Table 8: ERRANT error type analysis for mBART fine-tuned on 2.5K sample of Falko-MERLIN plus Wiki-Comments

Category	P	R	F0.5
ADJ	0.2121	0.1522	0.1966
ADJ:FORM	0.7914	0.6192	0.7497
ADP	0.4918	0.1802	0.3654
ADV	0.3333	0.0476	0.1515
ADV:FORM	0.8571	1.0	0.8824
AUX	0.5484	0.3676	0.4993
AUX:FORM	0.5283	0.3457	0.4778
CONJ	0.3	0.2143	0.2778
DET	0.4562	0.2607	0.3967
DET:FORM	0.7624	0.6914	0.747
MORPH	0.7297	0.3195	0.5806
NOUN	0.2051	0.1127	0.1762
NOUN:FORM	0.7692	0.5023	0.6953
ORTH	0.5312	0.4121	0.5022
OTHER	0.2962	0.1873	0.2653
PART	0.6286	0.4	0.5641
PNOUN	0.0	0.0	0.0
PRON	0.405	0.2121	0.3427
PRON:FORM	0.7556	0.4416	0.6615
PUNCT	0.645	0.5598	0.6259
SCONJ	0.4	0.1071	0.2586
SPELL	0.6951	0.2542	0.5161
VERB	0.507	0.2156	0.3991
VERB:FORM	0.88	0.1818	0.4977
WO	0.5795	0.3778	0.5236

Table 9: ERRANT error type analysis for mBART fine-tuned only on 5K sample of Falko-MERLIN

Category	P	R	F0.5
ADJ	0.2917	0.1522	0.2465
ADJ:FORM	0.7679	0.5397	0.708
ADP	0.6438	0.1411	0.376
ADV	0.2857	0.0238	0.0893
ADV:FORM	0.8333	0.8333	0.8333
AUX	0.5833	0.3027	0.4921
AUX:FORM	0.5778	0.321	0.4981
CONJ	0.5	0.1429	0.3333
DET	0.5217	0.1714	0.3704
DET:FORM	0.8198	0.6431	0.7771
MORPH	0.7719	0.2604	0.5542
NOUN	0.3095	0.1831	0.272
NOUN:FORM	0.7414	0.3927	0.6296
ORTH	0.5	0.3111	0.4459
OTHER	0.3665	0.177	0.3019
PART	0.6061	0.3636	0.5348
PNOUN	0.2222	0.2667	0.2299
PRON	0.4444	0.1212	0.2899
PRON:FORM	0.875	0.2727	0.6069
PUNCT	0.7107	0.5158	0.6608
SCONJ	0.5	0.0893	0.2604
SPELL	0.6812	0.3177	0.5544
VERB	0.6207	0.2156	0.4511
VERB:FORM	0.7826	0.1488	0.4225
WO	0.5616	0.3037	0.4801

Table 10: ERRANT error type analysis for mBART finetuned on 5K sample of Falko-MERLIN plus WikiComments

D Artifact Licenses

Falko is under CC BY 3.0, and MERLIN by CC BY-SA 4.0. RULEC-GEC needs a User Agreement form filled as per their code repo, the dataset is also CC BY-SA 4.0. The Kor-Learner corpus is only allowed to be used & distributed for non-commercial purposes. The Kor-Native corpus is also only allowed to be used for non-commercial purposes. Please check the code repo by Yoon et al. (2023): https://github.com/soyoung97/Standard_Korean_GEC/tree/main for more details. Wikipedia dump data are under CC BY-SA 3.0 besides some exceptions detailed here: https://dumps.wikimedia.org/legal.html. Our usage falls under their intended use. We release WikiComments under an MIT License.

E (M^2) Precision & Recall Graphs for RQ2

This section shows the Precision and Recall graphs for the corresponding F0.5 graphs in Section 5.2

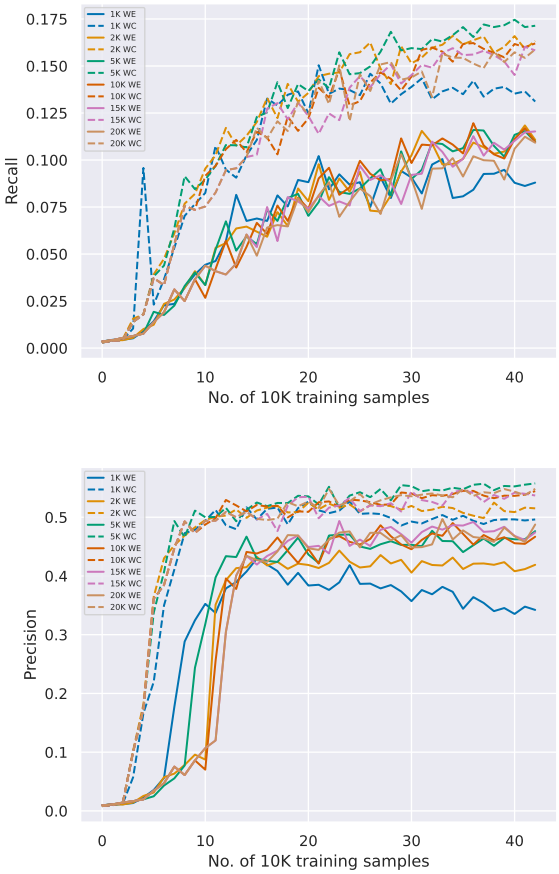

Figure 7: Comparison of German *WikiComments* and *WikiEdits* mBART runs evaluated on Falko-MERLIN

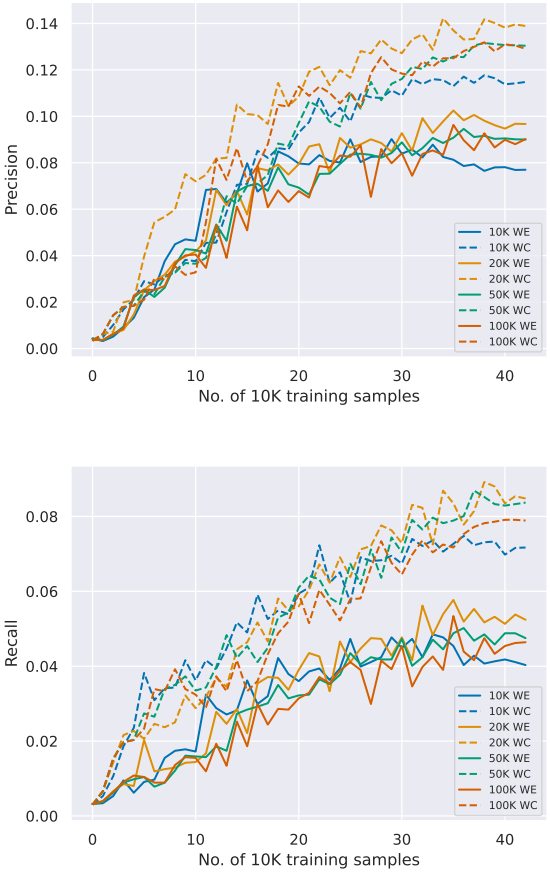

Figure 8: Comparison of Russian *WikiComments* and *WikiEdits* mBART runs evaluated on RULEC

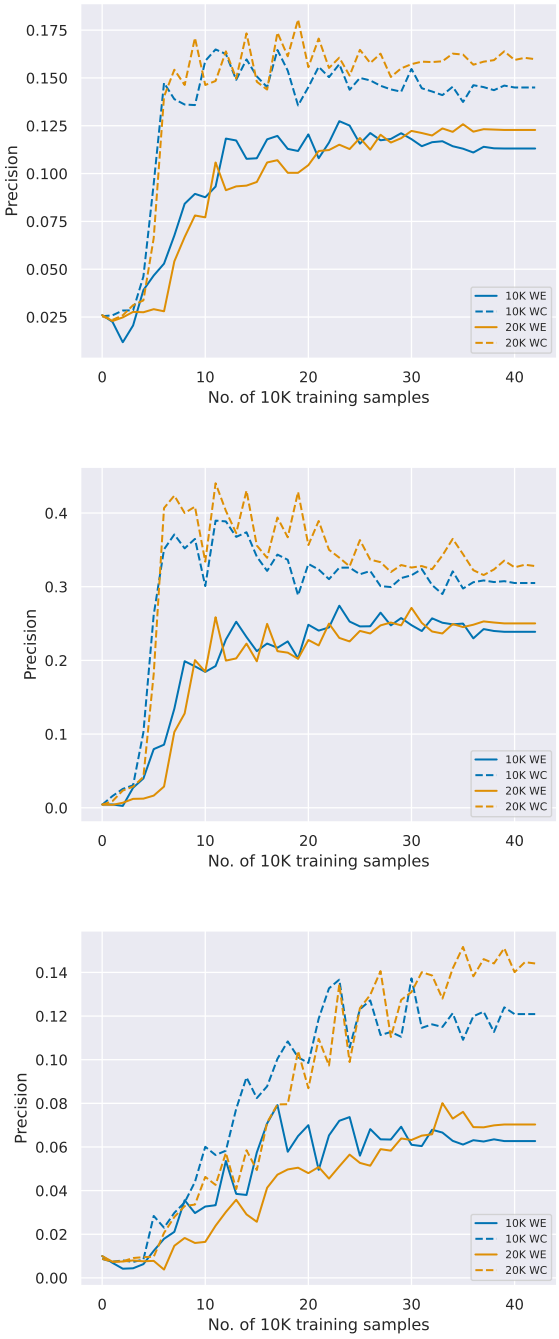

Figure 9: Comparison of Korean *WikiComments* and *WikiEdits* mBART runs evaluated on Kor-Learner, Kor-Union and Kor-Native

Figure 10: Comparison of Korean *WikiComments* and *WikiEdits* mBART runs evaluated on Kor-Learner, Kor-Union and Kor-Native