Improving Grammatical Error Correction with Multimodal Feature Integration

Tao Fang[1*] **Jinpeng Hu**[2*†] **Derek F. Wong**[1†] **Xiang Wan**[2]
Lidia S. Chao[1] **Tsung-Hui Chang**[2]
[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.taofang@gmail.com {derekfw,lidiasc}@um.edu.mo
[2]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong,
Shenzhen, Guangdong, China
jinpenghu@link.cuhk.edu.cn wanxiang@sribd.cn changtsunghui@cuhk.edu.cn

Abstract

Grammatical error correction (GEC) is a promising task aimed at correcting errors in a text. Many methods have been proposed to facilitate this task with remarkable results. However, most of them only focus on enhancing textual feature extraction without exploring the usage of other modalities' information (e.g., speech), which can also provide valuable knowledge to help the model detect grammatical errors. To shore up this deficiency, we propose a novel framework that integrates both speech and text features to enhance GEC. In detail, we create new multimodal GEC datasets for English and German by generating audio[1] from text using the advanced text-to-speech models. Subsequently, we extract acoustic and textual representations by a multimodal encoder that consists of a speech and a text encoder. A mixture-of-experts (MoE) layer is employed to selectively align representations from the two modalities, and then a dot attention mechanism is used to fuse them as final multimodal representations. Experimental results on CoNLL14, BEA19 English, and Falko-MERLIN German show that our multimodal GEC models achieve significant improvements over strong baselines and achieve a new state-of-the-art result on the Falko-MERLIN test set.

1 Introduction

Grammatical error correction (GEC) is one of the promising applications in natural language processing (NLP), aiming to correct sentences containing grammatical errors. GEC has attracted substantial attention in the past few decades owing to its importance in writing assistance for language learners (Rothe et al., 2021; Zhao and Wang, 2020; Qorib et al., 2022; Wan et al., 2020; Chollampatt and Ng, 2018; Tarnavskyi et al., 2022; Kaneko et al., 2020; Zhang et al., 2022a; Fang et al., 2023a; Zhang et al., 2023a; Fang et al., 2023b; Zhang et al., 2023b).

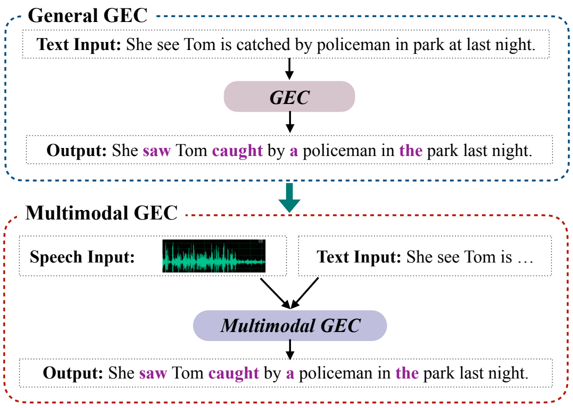

Figure 1: A comparison between general GEC and multimodal GEC tasks. The top is the general GEC system, which only relies on text modality, and the bottom is the proposed multimodal GEC task combining text and its corresponding speech.

In recent years, pre-trained Transformer-based models have proven effective in many NLP tasks (Hu et al., 2022a,b; Clinchant et al., 2019; Liu and Lapata, 2019; Hu et al., 2023b; Zhong et al., 2022; Liu et al., 2021; Li et al., 2022), including GEC (Gong et al., 2022; Li et al., 2023), because these models consist of multiple-layer multi-head attention and are trained with massive language data so that they are more powerful in feature extraction than other counterpart models. For example, Kaneko et al. (2020) first proposed to fine-tune BERT with the GEC corpus and then use the output of BERT as additional features to enhance GEC. Rothe et al. (2021) used the T5 structure (Raffel et al., 2020) to refine the GEC corpus (i.e., CLang8) and obtained promising results in GEC for different languages. Furthermore, Qorib et al. (2022); Tarnavskyi et al. (2022) employed binary classification or majority votes on span-level edits to ensemble multiple Transformer-based models.

Although these methods have achieved considerable improvements, they may focus on the bet-

*Equal Contribution
†Co-corresponding Author
[1]https://github.com/NLP2CT/MultimodalGEC

Findings of the Association for Computational Linguistics: ACL 2023, pages 9328–9344
July 9-14, 2023 ©2023 Association for Computational Linguistics

ter use of textual data while failing to take other modalities into consideration (e.g., speech). Many studies have shown that other modality data (e.g., speech) can effectively enhance feature extraction and thus promote model performance, such as risk forecasting (Sawhney et al., 2020), semantic matching (Huzaifah and Kukanov, 2022), etc. For example, Huzaifah and Kukanov (2022) studied a joint speech-text embedding space through a semantic matching objective and achieved better results in downstream tasks. Kim and Kang (2022) proposed to learn the cross-modality interaction between acoustic and textual information for emotion classification, which outperformed unimodal models. These works illustrate that audio signals can be regarded as complementary information and provide valuable features to promote text processing. Besides, intuitively, the audio with grammatical errors can be easily captured by the native speakers according to their spoken language experiences, which can implicate that speech should be effective in helping the model to distinguish whether the text contains ungrammatical elements.

Therefore, in this paper, we propose to integrate speech and text features to promote GEC, with an example shown in Figure 1. Firstly, owing to the lack of multimodal datasets for GEC, we adopt advanced text-to-speech (TTS) models to automatically generate audio for each instance in GEC datasets. Afterward, we extract acoustic and textual representations by a multimodal encoder that consists of pre-trained speech and text encoders. Furthermore, we propose to utilize an MoE layer to selectively align features from speech and text modalities, and then simple dot attention is applied to fuse them as final multimodal representations, which are then input to a pre-trained decoder to generate corrected sentences. Experimental results on English and German benchmarks illustrate the effectiveness of our proposed model, where our model achieves significant improvements over strong unimodal GEC baselines. Further analysis shows that our multimodal GEC model demonstrates significant improvements in most POS-based fine-grained error types, as well as in the major Operation-Level error types such as word substitutions, missing words, and unnecessary words.

The contributions are concluded as follows:

- To the best of our knowledge, this paper is the first to utilize a multimodal model to combine audio and text features to facilitate GEC.
- This paper constructs multimodal GEC datasets for English and German, where each sample in the dataset is a triple (ungrammatical text, audio, grammatical text).
- This paper proposes to use a mixture-of-experts module to dynamically align text and speech pairs for multimodal GEC.
- This paper reveals the gains and losses of incorporating speech modality into GEC on error types, providing clues for future research.

2 Data Construction

Owing to the lack of speech data in the GEC task, we need to construct multimodal GEC datasets for multimodal GEC tasks. Therefore, in this section, we give the details of dataset construction. Speech processing has achieved promising improvement over the past few decades, including converting sentences in the text into utterances (Ren et al., 2019; Qi et al., 2023). Therefore, we employ the advanced speech synthesis system to convert each piece of source side of GEC data (i.e., the ungrammatical side) into audio data to construct GEC multimodal data. As a result, each example in the GEC dataset is expanded into a triplet consisting of the ungrammatical sentence, the audio generated from the corresponding ungrammatical sentence, and the grammatical sentence.

2.1 English GEC Multimodal Data

For constructing the English GEC multimodal dataset, we adopt the FastSpeech2[2] text-to-speech model (Wang et al., 2021) to produce audio data from the source side of English GEC data. Specifically, to construct GEC multimodal training data, we convert the distilled English CLang8 GEC data (Rothe et al., 2021) into audio data. For constructing development and test sets, we select the widely-used CoNLL14 (Ng et al., 2014) and BEA19 (Bryant et al., 2019) English GEC benchmarks. For the CoNLL14 benchmark, the CoNLL13 (Ng et al., 2013) and the official-2014.combined.m2 version of CoNLL14 are used for constructing multimodal development and test sets, respectively. For the BEA19 benchmark, we use the BEA19 development and test sets to construct audio data.

[2] https://github.com/facebookresearch/fairseq/tree/main/examples/speech_synthesis

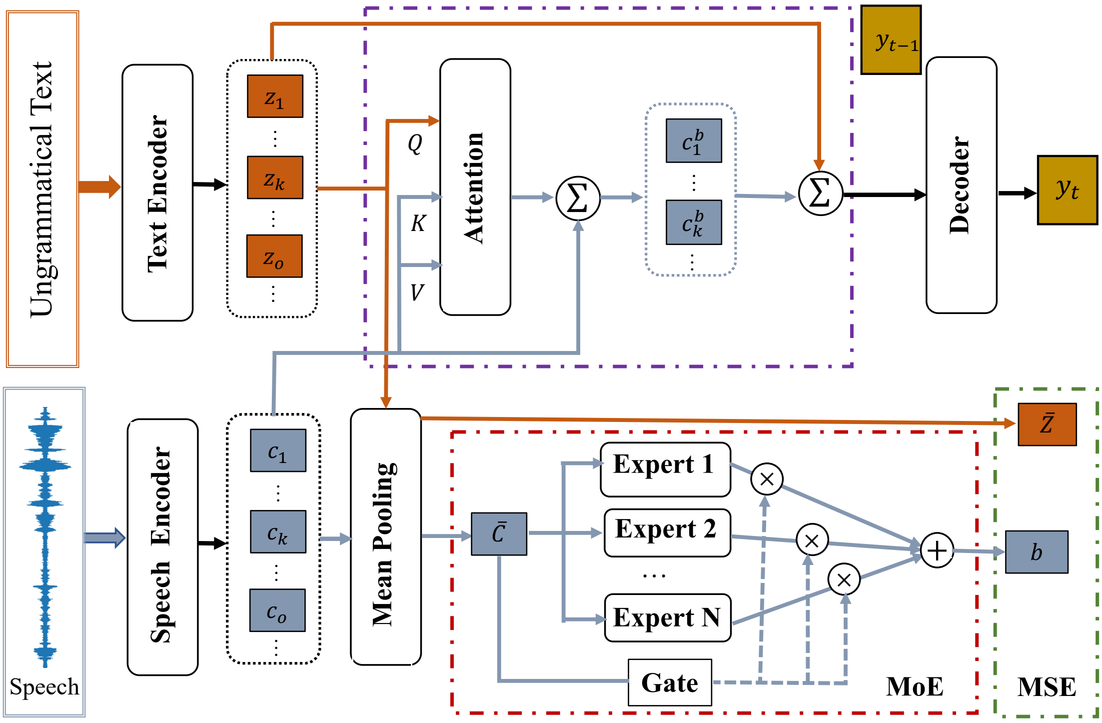

Figure 2: The overall framework of our proposed multimodal GEC model. The pre-trained text and speech encoders extract the features of the ungrammatical text and its corresponding acoustic. The red dotted box represents the MoE layer, which dynamically aligns audio and text. Dot attention fusion module, represented by the purple dotted box, is used to fuse the aligned textual and acoustic features as the final multimodal representations. The MSE objective (green dotted box) serves as a constraint during the feature fusion process.

LAN.	DATA	TRAIN (#Triples)	DEV (#Triples)	TEST (#Triples)
EN	CL8-EN	2.2M	-	-
	BEA19	-	4,384	4,477
	CONLL13	-	1,379	-
	CONLL14	-	-	1,312
DE	CL8-DE	110K	-	-
	FALKO-ME.	12.9K	2,503	2,337

Table 1: Statistics of the generated multimodal GEC datasets for English and German.

2.2 German GEC Multimodal Data

For building German multimodal GEC datasets, we employ gTTS (Google Text-to-Speech) toolkit[3] to generate audio data from the source side of German GEC training, development and test data. We build multimodal training data from German CLang8 and the official Falko-MERLIN (Boyd et al., 2014) training data. As for the multimodal development and test sets, we produce the audio data from Falko-MERLIN German validation and test sets.

[3] https://gtts.readthedocs.io/en/latest/

2.3 Data Processing

To prepare the text GEC datasets for audio generation, we first remove duplicate instances from the English CLang8 dataset, while keeping the other datasets unaltered. Additionally, we follow Katsumata and Komachi (2020) to use Moses script (Koehn et al., 2007) to detokenize GEC data for English and German. The statistics of the final multimodal datasets are shown in Table 1.

3 Method

3.1 Problem Definition

Existing approaches mainly utilize an encoder-decoder framework to address the GEC problem. In detail, the input is a sentence with grammatical errors $X = x_1, x_2, \cdots, x_N$, where N is the number of tokens, and the goal of this task is to correct the input sentence and generate a right one $Y = y_1, y_2, \cdots, y_L$, where L is the length of target sentence. Motivated by the success of multimodal in other tasks (Li et al., 2018; Sawhney et al., 2020), in this paper, we propose a novel multimodal GEC task and take a text-audio pair (X, S) as input (text and audio, respectively), aiming to integrate acoustic and textual features to enhance GEC. Therefore, the generation process for the multimodal GEC problem can be formulated as:

$$p(Y|X,S) = \prod_{t=1}^{L} p(y_t \mid y_1,\ldots,y_{t-1},X,S). \quad (1)$$

Moreover, we utilize the negative conditional log-likelihood of Y given the pair (X,S) to train the model:

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{L} \log p\left(y_t \mid y_1,...,y_{t-1},X,S;\theta\right), \quad (2)$$

where θ is the trainable parameters of the model. An overall structure of our proposed method is presented in Figure 2.

3.2 Multimodal Encoder

The multimodal encoder in our model consists of two main feature extractors: speech encoder and text encoder, respectively.

Speech Encoder We utilize a pre-trained Transformer-based model (e.g., wav2vec2 (Baevski et al., 2020)) as our speech encoder, which can learn powerful representations from speech audio and achieve promising results in many downstream tasks.

$$[\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_P] = f_{ae}(S), \quad (3)$$

where $\mathbf{c}$ is the features extracted from speech, f_{ae} refers to speech encoder and P is length of acoustic features.

Text Encoder We adopt a pre-trained model (e.g., T5 encoder (Raffel et al., 2020)) as our text encoder to capture textual features $\mathbf{z}$ from X:

$$[\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_N] = f_{te}(X), \quad (4)$$

where $\mathbf{z}_i$ is high dimensional vector for representing token x_i and f_{te} refers to the text encoder.

3.3 Multimodal Alignment and Fusion

On the one hand, it is intuitive that the speech should be semantically close to the corresponding text if they are in one pair since they actually represent similar meanings through different modalities. On the other hand, audio is used to provide complementary information instead of completely consistent information to help the model to better recognize and detect grammatical errors. As a result, we should allow some variance between features extracted from different modalities during multimodal alignment.

Therefore, we adopt a mixture-of-experts (MoE) to dynamically select semantically similar information from acoustic features, which is used to align with textual representation. The MoE layer in our model consists of M experts, denoted as $E_1, E_2, \cdots, E_M$, and each expert is a simple MLP with ReLU. Note that although these experts have identical structures, they have separate parameters instead of shared ones. We first obtain the overall representation of the speech S and text X by mean pooling, which can be formulated as:

$$\bar{\mathbf{c}} = Mean([\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_P]), \quad (5)$$

$$\bar{\mathbf{z}} = Mean([\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_N]), \quad (6)$$

We utilize the MoE to further extract the features from $\bar{\mathbf{c}}$ that should be close to $\bar{\mathbf{z}}$. Specifically, the output of i-th expert is denoted as $E_i(\bar{\mathbf{c}})$ and we follow Shazeer et al. (2017) to generate a gate $G_i(\bar{\mathbf{c}})$ for each expert. The output of the MoE module can be written as:

$$\mathbf{b} = \sum_{i=1}^{M} G_i(\bar{\mathbf{c}})E_i(\bar{\mathbf{c}}), \quad (7)$$

where $\mathbf{b}$ should be the information that is semantically close to the text. We utilize a simple mean squared error (MSE) objective to constrain this process and align these textual and acoustic features, which can be formulated as:

$$\mathcal{L}_{mse} = MSE(\mathbf{b},\bar{\mathbf{z}}), \quad (8)$$

After dynamic alignment between audio and text, we utilize dot attention to fuse these two features. In detail, we first compute the attention weight with the softmax function:

$$\mathbf{a}_i = \mathrm{Softmax}(\mathbf{z}_i\mathbf{c}^{\mathrm{T}}). \quad (9)$$

Herein, $\mathbf{a}_i$ can be viewed as a probability distribution and used to produce a weighted sum over the visual patch representations:

$$\mathbf{z}_i^c = \sum_{k=1}^{P} a_{i,k}\mathbf{c}_k. \quad (10)$$

Finally, we sum the $\mathbf{z}^c$ and $\mathbf{z}$ as final multimodal representation $\mathbf{h}$.

3.4 Decoder

The multimodal representation $\mathbf{h}$ is input to the pre-trained decoder (e.g., T5) to generate the correct sequence:

$$y_t = f_{de}(\mathbf{h}, y_1,\cdots,y_{t-1}) \quad (11)$$

This process is repeated until the complete sentence is obtained.

As for training, the final objective is the linear combination of losses from the sequence generation and multimodal alignment:

$$\mathcal{L} = \mathcal{L}_{ge} + \lambda \mathcal{L}_{mse}, \tag{12}$$

where $\mathcal{L}_{ge}$ is the basic sequence-to-sequence loss and λ is the weight to control the MSE loss.

4 Experimental Settings and Results

4.1 Data and Evaluation

The multimodal GEC data used for training is presented in Table 1 in section 2. With respect to English, we follow Rothe et al. (2021) and use only the English CLang8 multimodal data for training as they reported that further fine-tuning on high-quality English datasets, such as FCE v2.1 (Yannakoudakis et al., 2011) and W&I (Yannakoudakis et al., 2018), led to a drop in performance. For validation, we use the CoNLL13 multimodal data and the BEA19 multimodal development data when testing on the CoNLL14 and BEA19 English test sets, respectively. In terms of German, we first train our models on the German CLang8 multimodal data as Rothe et al. (2021), and then fine-tune the models on the official Falko-MERLIN German multimodal training data. For the development and test data, we use the official Falko-MERLIN German benchmark. Additionally, to establish a stronger baseline, we follow Katsumata and Komachi (2020) to use the same 10M synthetic data (Náplava and Straka, 2019)[4] to pre-train T5/mT5-Large model for English and German.

For evaluation, we use the M2 scorer (Dahlmeier and Ng, 2012) to evaluate the model performance on the CoNLL14 English test and the official Falko-MERLIN German benchmark. The BEA19 English test is evaluated by ERRANT (Bryant et al., 2017). We employ the T-test method to test the significance of the results, except for the BEA19 English test, which is a blind test set.

4.2 Implementation Details and Training

In our experiments, we adopt `Huggingface`[5] library to build our multimodal GEC model. Specifically, for the basic experiments, we utilize T5-Large (Raffel et al., 2020) and mT5-Large (Xue et al., 2020) as our text backbone models (including both text encoder and decoder), with the former being used for English and the latter for German. We follow their default setting, which uses 24 layers of self-attention with 16 heads. For the experiments with stronger baselines, we use our T5-Large and mT5-Large models fine-tuned on 10M synthetic data as text backbone models for English and German, respectively. The details of the training settings can be found in Appendix A.1. For the speech encoder, we adopt Hubert Large pre-trained model (Hsu et al., 2021) to extract features for English audio and wav2vec2-xls-r-300m pre-trained model (Babu et al., 2021) for German speech. We also follow the default settings for these speech models. As for training, we utilize Adafactor (Shazeer and Stern, 2018) to optimize all trainable parameters in our model. We set the number of experts to 6. The weight hyper-parameter λ is set to 0.1 for both English and German experiments. The other settings for training the multimodal GEC models are reported in Appendix A.2.

4.3 Baselines

To explore the effect of the proposed multimodal model for GEC, we compare our model with the following baselines:

- **LRGEC** (Náplava and Straka, 2019): it pre-trains a Transformer seq2seq model on synthetic data and then fine-tunes on authentic data.
- **TAGGEC** (Stahlberg and Kumar, 2021): the model improves GEC performance by data augmentation (e.g., generating synthetic data with the guidance of error type tags).
- **GECTOR** (Omelianchuk et al., 2020), **TMTC** (Lai et al., 2022), **EKDGEC** (Tarnavskyi et al., 2022): these models utilize the sequence tagging approach to improve GEC performance with multiple stage training, where they firstly pre-train on errorful-only sentences and further fine-tune on a high-quality dataset.
- **SADGEC** (Sun et al., 2021), **gT5 XXL** and **T5/MT5 LARGE/XXL** (Rothe et al., 2021): these GEC models borrow knowledge from pre-trained language models, where SADGEC is based on the BART (Lewis et al., 2020) pre-trained model, gT5 XXL is a large teacher model for distilling Lang8 data, which is first pre-trained from scratch on a large amount of synthetic data followed by fine-tuning on high-quality data. T5/MT5 LARGE/XXL adopt

[4] https://github.com/ufal/low-resource-gec-wnut2019/tree/master/data

[5] https://github.com/huggingface/transformers

SYSTEM	CoNLL14			BEA19 (TEST)		
	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$
LRGEC (Náplava and Straka, 2019)	-	-	63.4	-	-	69.0
GECToR (Omelianchuk et al., 2020)	77.5	40.1	65.3	79.2	53.9	72.4
TagGEC (Stahlberg and Kumar, 2021)	72.8	49.5	66.6	72.1	64.4	70.4
SADGEC (Sun et al., 2021)	71.0	52.8	66.4	-	-	72.9
TMTC (Lai et al., 2022)	77.8	41.8	66.4	81.3	51.6	72.9
EKDGEC (Tarnavskyi et al., 2022)	74.4	41.1	64.0	80.7	53.4	73.2
T5 LARGE (Rothe et al., 2021)	-	-	66.0	-	-	72.1
T5 XXL (Rothe et al., 2021)	-	-	68.8	-	-	**75.9**
gT5 XXL (Rothe et al., 2021)	-	-	65.7	-	-	69.8
OURS (T5 LARGE)	73.6	52.7	68.2	75.5	67.9	73.9
OURS (PRET5 LARGE)	75.0	53.2	**69.3**	77.1	66.7	74.8

Table 2: Results on the CoNLL14 and BEA19 English GEC test sets. Our multimodal GEC systems (**OURS**) are fine-tuned on the same CLang8 English data as T5 LARGE/XXL (Rothe et al., 2021). **PRET5 LARGE** means using the same 10M synthetic data as LRGEC (Náplava and Straka, 2019) to pre-train T5 large model, which can report a much stronger baseline when fine-tuned on CLang8 data (see Table 4). Notably, all the reported comparison results are a single model without ensembling. **Bold** values indicate the best $F_{0.5}$ scores.

SYSTEM	DATA	FALKO-ME.		
		Pre.	Rec.	$F_{0.5}$
LRGEC	offic.	78.2	59.9	73.7
MT5 LARGE	cl8	-	-	70.1
MT5 XXL	cl8	-	-	74.8
gT5 XXL	offic.	-	-	76.0
OURS (MT5)	cl8	76.1	59.8	72.1
	+offic.	77.2	65.4	$74.5^{\dagger}$
OURS (PMT5)	cl8	77.6	63.0	74.2
	+offic.	78.5	68.4	$\mathbf{76.3}^{\dagger}$

Table 3: Results on Falko-MERLIN GEC test set. **MT5** refers to mT5 large model, **PMT5** means using the same 10M German synthetic data as LRGEC to pre-train mT5 large model. offic. refers to the official Falko-MERLIN GEC training data, cl8 is the distilled German CLang8 data. Using the official data to fine-tune the models on cl8 can significantly improve performance ($^{\dagger}p < 0.01$).

T5/mT5 as the backbone structure and fine-tune on the corresponding distilled CLang8 data for GEC tasks in different languages.

4.4 Experimental Results

Results on English dataset To illustrate the effectiveness of our proposed model, we compare our model with existing studies with the results reported in Table 2. We obtain several observations from the results. First, the comparison between **OURS** and other baselines illustrate the effectiveness of our design in the GEC task, where our model achieves much better performance even though these competitors utilize many ways (e.g., data augmentation) to enhance feature extraction in GEC. The reason might be that compared to pure textual information, audio can provide complementary information to help the model better grasp the grammatical error in the sentence, and our model can selectively align these features from speech and text by the MoE module. It is easy to follow that a native speaker can distinguish whether the audio is grammatically correct. Second, compared to the sequence tagging method (e.g., GECToR), sequence-to-sequence based models (e.g., T5-LARGE) perform better in recall score but are inept at precision. Especially it is found that the strength of our proposed model lies in its high recall compared to other baselines. Third, continuing training T5-Large on 10M synthetic data can further improve the model performance, illustrating that synthetic data can alleviate the gap between GEC data and pre-training corpus. Appendix A.3 shows some examples generated by the unimodal and multimodal GEC models.

Results on German dataset To further demonstrate the validity of our model, we also conduct experiments on the German dataset, with the results reported in 3. We can obtain similar trends as in English GEC, where our proposed mode out-

Model	CoNLL14	Falko-ME.
	P/ R/ $F_{0.5}$	P/ R/ $F_{0.5}$
Ours ((m)T5)	73.6/ 52.7/ 68.2†	76.1/ 59.8/ 72.1†
−MoE	73.0/ 52.9/ 67.9	75.5/ 59.8/ 71.7
−Speech Enc.	72.2/ 51.4/ 66.8	75.7/ 56.5/ 70.9
Ours (P(m)T5)	75.0/ 53.2/ 69.3†	77.6/ 63.0/ 74.2†
−MoE	74.8/ 52.6/ 69.0	77.6/ 62.8/ 74.1
−Speech Enc.	73.5/ 53.7/ 68.5	77.3/ 62.3/ 73.8

Table 4: Ablation results of our proposed method on the CoNLL14 English and the Falko-MERLIN German tests, which were trained on CLang8 data. Statistically significant improvements over "−Speech Enc." model, as indicated by P_value, $^\dagger p < 0.01$

performs other baselines and achieves a superior $F_{0.5}$ score. Especially, by further fine-tuning the models on the official data, we achieve a new state-of-the-art result (i.e., 76.3 $F_{0.5}$). This result further demonstrates that audio can provide valuable benefits in GEC tasks regardless of language type. Additionally, even though the German dataset is much smaller compared to the English dataset, our model still achieves significant improvements, which highlights its effectiveness in low-resource settings.

5 Analyses

5.1 Ablation Study

To explore the effectiveness of our proposed method, we conduct the ablation studies with the following settings: a) removing the MoE layer (−MoE) and retaining the dot attention module to fuse acoustic and textual features. b) removing the speech encoder (−Speech Enc.), which degenerates our multimodal GEC model into a text-only unimodal GEC model. As shown in Table 4, when we remove the MoE layer, the results of the multimodal GEC model show a decrease in all settings, demonstrating the validity of MoE in the multimodal feature fusion. Moreover, if we discard the speech encoder, the results of the reverted text-only unimodal GEC baseline models are significantly lower than the multimodal model for both English and German, which illustrates the effectiveness of our proposed multimodal GEC models.

5.2 Error Type Performance

To investigate the ability of GEC systems to correct different error types, we used the ERRANT toolkit (Bryant et al., 2017) to analyze the evaluation results on the CoNLL14 test set with respect to both POS-based fine-grained error types and Operation-Level error types.

Fine-grained Error Types Figure 3 shows the performance of the POS-based fine-grained error types. We can observe that while multimodal GEC is inferior to text-only unimodal GEC systems in certain error types (i.e., PUNCT, ADV, CONJ, and PREP), our model obtains better results in most types of errors, including ADJ, NOUN, NOUM: NUM, PRON, VERB, VERB: TENSE, DET, MORPH, ORTH, and PART, which further confirms the effectiveness of multimodal feature integration in the GEC task. In fact, adverb and conjunction error types account for a relatively small percentage of all grammatical errors (not more than 1.6%). In other words, multimodal GEC can improve the performance of common errors in GEC and thus bring considerable improvements overall.

Operation-Level Error Types We evaluate the performance of Operation-Level error types using the ERRANT toolkit, which categorizes them into three categories: **R**eplacement, **M**issing, **U**nnecessary. Considering that word order (**WO**) is a sub-type of Replacement, which is different from other types of errors, we manually separate into a separate category. As shown in Table 5, compared to text-only unimodal GEC baseline models, our multimodal GEC models are better at correcting the major operation-level error types, such as word substitutions (64.3%), missing words (17.9%), and unnecessary words (17.0%), demonstrating that the corresponding speech information is beneficial to GEC. However, the multimodal GEC model does not perform well in correcting word order, even if it is a minor issue (0.8%). We hypothesize that correcting word order requires sentence structure information (Zhang et al., 2022b), but the speech may not provide such information to GEC models.

6 Related work

Grammatical Error Correction (GEC) is the task of automatically identifying and correcting grammatical errors in a text (Ng et al., 2013). Previous research in this field has primarily focused on strengthening the representations of text data through data augmentation techniques, such as using the back-translation method (Sennrich et al., 2016) for the GEC task (Kasewa et al., 2018; Xie et al., 2018; Kiyono et al., 2019), and injecting noise with specific rules into grammatical sentences (Lichtarge et al., 2019; Zhao et al., 2019; Xu et al., 2019; Stahlberg and Kumar, 2021). More recently, pre-trained language models (PLMs) have

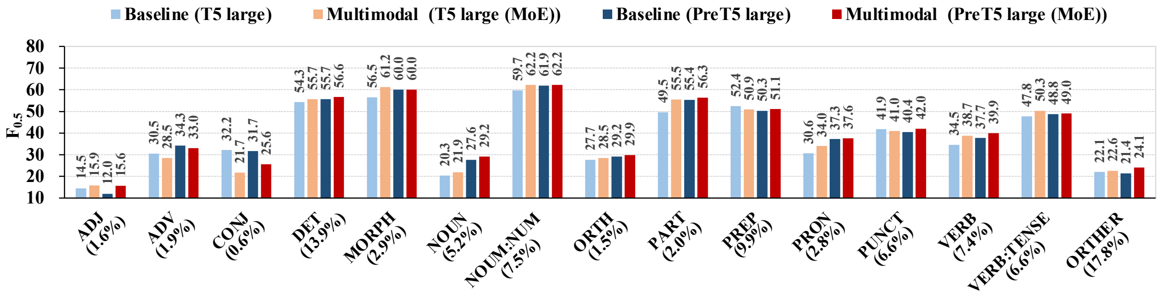

Figure 3: $F_{0.5}$ scores on a selection of fine-grained error types on the CoNLL14 test set, with the percentages in parentheses indicating the proportion of each error type. Overall, the results show that integrating speech modality information into text-only GEC can significantly improve the performance on most fine-grained error types.

Method	R (64.3%)			M (17.9%)			U (17.0%)			WO (0.8%)		
	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$
T5 (Base.)	50.9	37.1	47.4	42.5	38.1	41.5	56.0	35.9	50.4	35.7	46.2	**37.4**
T5 (MultiM.)	52.5	36.5	**48.3**	45.7	37.3	**43.7**	58.9	34.2	**51.5**	31.4	35.5	32.1
PreT5 (Base.)	52.3	37.9	48.6	45.3	37.8	43.6	57.8	35.1	51.2	37.1	50.9	**39.2**
PreT5 (MultiM.)	53.0	37.0	**48.8**	47.2	37.2	**44.8**	59.8	35.7	**52.7**	37.0	43.2	38.1

Table 5: Performance by Operation-Level error types on the CoNLL14 test set for text-only unimodal and fused speech and text multimodal GEC models. The percentages in parentheses represent the proportion of operation-level error types. Results in **bold** indicate the best $F_{0.5}$ scores. The multimodal GEC models demonstrate improved accuracy for the major operation-level error types, such as Substitution, Insertion, and Deletion.

been demonstrated to be effective in improving the performance of GEC tasks. Studies such as Choe et al. (2019) have leveraged sequential transfer learning to adapt pre-trained Transformer models to the GEC domain. Kaneko et al. (2020) initialized an encoder-decoder GEC model with pre-trained BERT weights to enhance GEC performance. Katsumata and Komachi (2020) utilized the pre-trained BART model as a generic pre-trained encoder-decoder model for GEC, and Rothe et al. (2021) adopted a pre-trained T5 model to distill GEC corpus and used the pre-trained structure as part of the network for distilled GEC training, achieving promising results. However, to date, no previous work has attempted to incorporate multimodal information (e.g., speech modality) into the GEC task. Our work is the first to explore the use of multimodal information for GEC.

Multimodal Many studies have demonstrated the potential of incorporating multimodal information in improving the performance of single-modal tasks in the NLP domain. For example, Schifanella et al. (2014) and Cai et al. (2019) integrated image modality into the Twitter sarcasm detection task and found that incorporating image information can enhance the performance of this text-only task. Hu et al. (2023a) proposed to integrate radiology images and textual findings to improve impression generation. Additionally, Zheng et al. (2021) fused acoustic and text encoding to jointly learn a unified representation, thereby improving speech-to-text translation tasks. Li et al. (2017) demonstrated that fusing speech modality can enhance the readability of text summarization tasks. Huzaifah and Kukanov (2022) studied a joint speech-text embedding space through a semantic matching objective, achieving improved results in downstream tasks. Kim and Kang (2022) proposed a method for learning the cross-modality interaction between acoustic and textual information, which outperformed the unimodal models in emotion classification. In this work, we are the first to attempt to fuse acoustic and text to improve the GEC task.

7 Conclusion

This paper presents a novel approach to the task of multimodal GEC that integrates speech and text features to improve grammatical error correction. Due to the scarcity of speech data in GEC, we expand the original GEC data to create new multimodal GEC datasets for English and German, where each sample in our datasets is a triple (grammatically incorrect text, audio, and corrected text). Our approach utilizes a speech and text encoder to extract acoustic and textual features from the speech and input text, respectively. Then, we employ an MoE approach to selectively extract audio features that align with the textual features and use a dot attention layer to fuse the features from different modalities as the final representation. This fused representation is input to the decoder to generate the corrected sentence. Our experimental results on widely-used benchmarks demonstrate the effectiveness of our proposed model, achieving significant improvements compared to existing studies.

Limitations

Our proposed multimodal Grammatical Error Correction (GEC) model is based on a Seq2Seq generative framework, which utilizes different encoders to extract information from each modality, and then fuses them to provide input to an autoregressive decoder. However, in this work, we did not explore the use of a sequence tagging framework, which may be a consideration for future research, as it has the advantage of faster decoding speed. Additionally, this study focuses on the use of audio representations of the source-side of GEC data, rather than the target-side, to construct multimodal GEC data. Our further analysis concludes that our proposed multimodal GEC model has limitations in correcting certain minor error types (e.g., ADV, CONJ, PUNCT, and word order) when compared to text-only GEC models.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), and the Multi-year Research Grant from the University of Macau (Grant No. MYRG2020-00054-FST), and the Shenzhen Science and Technology Program (JCYJ20220818103001002), and the Guangdong Provincial Key Laboratory of Big Data Computing, the Chinese University of Hong Kong, Shenzhen. This work was performed in part at SICC which is supported by SKL-IOTSC, and HPCC supported by ICTO of the University of Macau.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. 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 (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeoil Yoon. 2019. A neural grammatical error correction system built on better pre-training and sequential transfer learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–227, Florence, Italy. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multi-layer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Tao Fang, Xuebo Liu, Derek F. Wong, Runzhe Zhan, Liang Ding, Lidia S. Chao, Dacheng Tao, and Min Zhang. 2023a. Transgec: Improving grammatical error correction with translationese. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

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

Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. 2022. Revisiting grammatical error correction evaluation and beyond. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6891–6902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jinpeng Hu, Zhihong Chen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2023a. Improving radiology summarization with radiograph and anatomy prompts. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Jinpeng Hu, DanDan Guo, Yang Liu, Zhuo Li, Zhihong Chen, Xiang Wan, and Tsung-Hui Chang. 2023b. A Simple Yet Effective Subsequence-Enhanced Approach for Cross-Domain NER. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022a. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4677–4688.

Jinpeng Hu, Yaling Shen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2022b. Hero-gang neural model for named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1924–1936.

Muhammad Huzaifah and Ivan Kukanov. 2022. Analysis of joint speech-text embeddings for semantic matching. *arXiv preprint arXiv:2204.01235*.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online. Association for Computational Linguistics.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4977–4983, Brussels, Belgium. Association for Computational Linguistics.

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*, pages 827–832, Suzhou, China. Association for Computational Linguistics.

Donghwa Kim and Pilsung Kang. 2022. Cross-modal distillation with audio–text fusion for fine-grained emotion classification using bert and wav2vec 2.0. *Neurocomputing*, 506:168–183.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Shaopeng Lai, Qingyu Zhou, Jiali Zeng, Zhongli Li, Chao Li, Yunbo Cao, and Jinsong Su. 2022. Type-driven multi-turn corrections for grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3225–3236, Dublin, Ireland. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Bei Li, Quan Du, Tao Zhou, Yi Jing, Shuhan Zhou, Xin Zeng, Tong Xiao, JingBo Zhu, Xuebo Liu, and Min Zhang. 2022. ODE transformer: An ordinary differential equation-inspired model for sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8335–8351, Dublin, Ireland. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4152–4158.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.

Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F. Wong, Yang Gao, Heyan Huang, and Min Zhang. 2023. Templategec: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. 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, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, and Zhaopeng Tu. 2021. Understanding and improving encoder layer fusion in sequence-to-sequence learning. In *International Conference on Learning Representations*.

Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.

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 2019)*, pages 346–356, Hong Kong, China. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Heli Qi, Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2023. Speechain: A speech toolkit for large-scale machine speech chain. In *arXiv*.

Muhammad Qorib, Seung-Hoon Na, and Hwee Tou Ng. 2022. Frustratingly easy system combination for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974, Seattle, United States. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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 (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM international conference on multimedia*, pages 456–465.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2014. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Xin Sun, Tao Ge, Furu Wei, and Houfeng Wang. 2021. Instantaneous grammatical error correction with shallow aggressive decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5937–5947, Online. Association for Computational Linguistics.

Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3842–3852, Dublin, Ireland. Association for Computational Linguistics.

Zhaohong Wan, Xiaojun Wan, and Wenguang Wang. 2020. Improving grammatical error correction with data augmentation by editing latent representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2202–2212, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. 2021. fairseq sˆ2: A scalable and integrable speech synthesis toolkit. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 143–152, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 619–628, New Orleans, Louisiana. Association for Computational Linguistics.

Shuyao Xu, Jiehao Zhang, Jin Chen, and Long Qin. 2019. Erroneous data generation for grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 149–158, Florence, Italy. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *arXiv preprint arXiv:2010.11934*.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023a. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *arXiv preprint arXiv:2305.13225*.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022a. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States. Association for Computational Linguistics.

Yue Zhang, Bo Zhang, Haochen Jiang, Zhenghua Li, Chen Li, Fei Huang, and Min Zhang. 2023b. Nasgec: a multi-domain chinese grammatical error correction dataset from native speaker texts. *arXiv preprint arXiv:2305.16023*.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022b. Syngec: Syntax-enhanced grammatical error correction with a tailored gec-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2518–2531. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota. Association for Computational Linguistics.

Zewei Zhao and Houfeng Wang. 2020. Maskgec: Improving neural grammatical error correction via dynamic masking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1226–1233. Association for the Advancement of Artificial Intelligence (AAAI).

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *International Conference on Machine Learning*, pages 12736–12746. PMLR.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

A Appendix

A.1 Pre-training Settings for T5/mT5-Large Model

The settings of hyper-parameters for pre-training T5/mT5-Large models for English and German are listed in Table 6.

CONFIG.	ENGLISH MODEL	GERMAN MODEL
Model Arch.	T5-Large	mT5-Large
Optimizer	Adafactor	Adafactor
Learning Rate	0.0008	0.0007
Batch Size	24	16
Update Freq.	128	64
GPUs	2 (A100)	2 (A100)

Table 6: Hyper-parameters for pre-training T5/mT5-Large models on 10M synthetic GEC data for English and German. Model Arch. refers to model architecture, Update Freq. means gradient accumulation steps.

A.2 Settings of Training Multimodal GEC Models

Table 7 presents the settings of hyper-parameters for training English and German multimodal GEC models.

CONFIG.	ENGLISH MULTIM.	GERMAN MULTIM.
	Stage-I	
Text backbone	T5-Large	mT5-Large
Speech Encoder	Hubert-Large	wav2vec2-xls-r-300m
Optimizer	Adafactor	Adafactor
Learning Rate	0.0001	0.0002
Batch Size	16	8
Update Freq.	16	16
Num. of Experts	6	6
K	2	2
λ	0.1	0.1
	Stage-II	
Optimizer	-	Adafactor
Learning Rate	-	0.0001
Batch Size	-	8
Update Freq.	-	2
Num. of Experts	-	6
K	-	2
λ	-	0.1
	Generation	
Beam size	5	5
Max input length	128	128

Table 7: Hyper-parameters for training English and German multimodal GEC models.

A.3 Case Study

Table 8 shows some examples generated by the text-only unimodal GEC model and multimodal GEC model. Our multimodal GEC model is better at correcting common error types (e.g. VERB) while exhibiting inferior performance in correcting word order errors.

Src	A couple did not have a child after their marriage for a long time**, their** parents were anxious about that and asked them to go to hospital to check what **was the problem**.
Ref.	A couple did not have a child after their marriage for a long time**. Their** parents were anxious about that and asked them to go to hospital to check what **the problem was**.
T5 (Base.)	A couple did not have a child after their marriage for a long time**. Their** parents were anxious about that and asked them to go to hospital to check what **the problem was**.
T5 (MoE)	A couple did not have a child after their marriage for a long time**. Their** parents were anxious about that and asked them to go to hospital to check what **was the problem**.
Src	Spouses usually have very close relationships, if person A **tell** his family that he has this gene, his uncle C knows and tells his wife D that he **needed** to run a test because his **cousine** has this disease.
Ref.	Spouses usually have very close relationships. If person A **tells** his family that he has this gene, his uncle C knows and tells his wife D that he **needs** to run a test because his **cousin** has this disease .
T5 (Base.)	Spouses usually have very close relationships. If person A **tells** his family that he has this gene, his uncle C knows and tells his wife D that he **needed** to run a test because his **cousin** has this disease.
T5 (MoE)	Spouses usually have very close relationships. If person A **tells** his family that he has this gene, his uncle C knows and tells his wife D that he **needs** to run a test because his **cousin** has this disease.

Table 8: Examples of the outputs generated by the unimodal/multimodal GEC model. **Src** refers to the ungrammatical sentence, and **Ref.** is the grammatical sentence. **T5 (Base.)** refers to the outputs of the unimodal GEC model. **T5 (MoE)** refers to the outputs of our multimodal GEC baseline model. The words with the color **red** are the ungrammatical parts and the **blue** indicates the corrected version.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section

☐ A2. Did you discuss any potential risks of your work?
Not applicable. There are no potential risks associated with this paper because all tasks we used are public ones that have been verified for years.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract section, and Section 1

☑ A4. Have you used AI writing assistants when working on this paper?
We use ChatGPT AI writing assistants to check some spelling errors and polish some sentences of our work (i.e., sections 4.1, 4.2, and 7)

B ☑ Did you use or create scientific artifacts?

Section 2, Section 4, and Section 4.1

☑ B1. Did you cite the creators of artifacts you used?
section 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All datasets and models we used here are public without restriction for research purposes.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
All datasets and models we used here are public without restriction for research purposes.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The datasets we used in our paper do not have such issues according to the claims in the original paper.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The datasets we used in our paper do not have such issues according to the claims in the original paper.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2, and Section 4.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 2, Section 4, Section 4.1, Section 4.4, Section 5.1, Appendix A.1, A.2, Table 3, and Table 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.1, A.2

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4, Appendix A.1, A.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.4, Section 5.1, Table 3, Table 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 2, Section 4, and Appendix A.1, A.2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.