ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890578

EPIC: Error Pattern Informed Correction for Classroom ASR with Limited Labeled Data

Linzhao Jia[1,2,3], Han Sun[1,2,3], Yuang Wei[1,2,3], Changyong Qi[1,2,3], Xiaozhe Yang[1,2,4]*
[1]*Lab of Artificial Intelligence for Education, East China Normal University, Shanghai, China*
[2]*Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai, China*
[3]*School of Computer Science and Technology, East China Normal University, Shanghai, China*
[4]*Institute of Curriculum and Instruction & Classroom Analysis Lab, East China Normal University, Shanghai, China*

***Abstract*—Automatic speech recognition (ASR) systems have a wide range of applications in classroom analysis. However, due to the unique structure of classroom dialogue, existing ASR systems often struggle to accurately recognize and organize spoken utterances, creating significant challenges for downstream tasks in educational dialogue analysis. To address this issue, we propose EPIC, a post-processing framework for classroom ASR error correction. We begin by extracting error patterns to gain a deeper understanding of the distribution of ASR errors. Next, we utilize large language models (LLMs) to reconstruct contextual information based on these error patterns, offering a viable solution for error correction with limited labeled data. Finally, after fine-tuning an error correction model, we implement a candidate selection process to identify the most appropriate hypothesis for each context. Extensive experiments with our proposed method demonstrate substantial improvements in word error rate (WER) and overall robustness in ASR error correction, enabling more reliable analysis of educational dialogues and offering deeper insights for educational research.**

***Index Terms*—ASR Error Correction, Classroom Dialogue, Large Language Models, Limited Labeled Data.**

I. Introduction

Automatic speech recognition (ASR) systems are extensively used to convert audio signals into word sequences for various downstream tasks [1], [2]. In educational settings, accurate transcription of classroom dialogues is essential for analyzing teacher-student interactions. However, classroom transcripts are often noisy, with frequent misspellings, missing words, and poorly formed syntax [3]–[5]. To address these challenges, ASR post-processing aims to rectify grammatical errors or omissions in sentences. Transformer-based ASR error correction methods [6]–[8] learn to map ASR outputs to corresponding ground-truth transcriptions. However, these methods may struggle with robustness when faced with the diverse structures of classroom dialogues and the limitations of a small training corpus. Another common approach is to rescore the N-best hypotheses generated by ASR systems, with the goal of predicting true transcription by evaluating these hypotheses [9], [10]. Meanwhile, the use of pre-trained language model for rescoring has shown promise, as they benefit from rich contextual information compared to rewrite the single best hypothesis [11], [12]. However, these hypotheses tend to be highly similar since they are derived from the same audio segment. A more recent and promising approach involves leveraging large language models (LLMs) through prompt-based methods to correct ASR errors, offering an alternative to traditional rescore techniques [13]–[18]. To address the problem of limited labeled data in ASR error correction, some works [19], [20] introduce noise into the training process to enhance the model's ability to handle diverse and unpredictable input data. While effective in some domains, these methods may not be directly applicable to the unique challenges presented in classroom settings. Additionally, classroom dialogues often exhibit a strong dependency on contextual information, which differs from the standard ASR error correction tasks that typically focus on isolated sentences. As shown in Table I, common ASR errors in classroom dialogue settings fall into four categories, with each error type paired with its corrected sentence on the right.

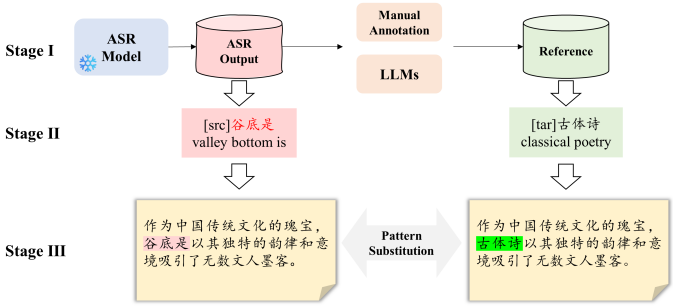

Fig. 1. Synthetic data generation through contextual augmentation.

To address these challenges, we propose the **EPIC** model, which utilizes LLMs and manually refinement of multiple sets of dialogue pairs to produce clean, high-quality ground truth text. A key component of our method is the error pattern extraction module, which can accurately identify and focus on error distribution. Then we employ LLMs to reconstruct the contextual information, complete the data augmentation process, and solve the problem of limited labeled data. Finally, we introduce a contextual similarity mechanism to select the best hypothesis for the current context, ensuring that the final corrected output is as contextually accurate as possible. This candidate selection process further improves the robustness of the correction system. Extensive experiments have shown

* Corresponding author.

TABLE I
TYPICAL ASR ERRORS FROM CLASSROOM DIALOGUE DATASET.

Type	ASR Hypothesis	Gold Reference
Similar sounding	谷底是 [valley bottom is]	古体诗 [classical poetry]
Add	5元5言句 [5 yuan 5-word sentence]	五言句 [five-word sentence]
Delete	望月⋆什么意思? [What does "Wang Yue"(looking at the moon) mean?]	望岳这首诗是什么意思? [What does this poem "Wang Yue"(Looking at the Mountain) mean?]
Domain-specific entities	我们能不能够出5月的一些独特的景观，或者还与其他阅不同的地方? [Can we mention some unique sceneries of May, or any different aspects compared to other Yue?]	我们能不能说出五岳的一些独特的景观，或者它与其他岳有什么不同? [Can we name some unique landscapes of the Five Sacred Mountains, or what sets them apart from other mountains?]

that EPIC delivers significant improvements across evaluation metrics and proves effective in low-resource scenarios.

II. PROPOSED MODELS

In this section, we present our data synthesis method, which involves three key steps. First, we leverage LLMs and manual annotation to refine a small subset of the data (Section II-A). Next, we construct an error pattern pool (Section II-B) and proceed to synthesize a contextual corpus and perform substitutions to generate the corpus (Section II-C). Finally, we fine-tune a pre-trained language model (PLM) on this synthesized data to develop a robust ASR error correction model (Section II-D).

A. Stage I: Data annotation

Our dataset mainly consists of Mandarin dialogues, transcribed using WhisperX[1] [21] with `Whisper-large-v3`[2] as the backbone. This model has been fine-tuned for high accuracy in complex, multi-speaker environments and provide ASR output in the following format:

$$S = [\text{Timestamp}][\text{Speaker}](w_1, w_2, , , w_n) \quad (1)$$

where each timestamp indicates the start and end of an utterance, the speaker identifier differentiates participants in the dialogue, and the sequence represents the words within each spoken segment.

The ASR model generates raw transcription output from spoken input, which is initially refined by LLMs to correct errors using contextual information. If the corrected sentence remains misaligned with the natural context, human annotators review the original audio to manually resolve the errors. This hybrid approach, which starts with LLM-based refinement followed by manual review when necessary, streamlines the process, saving time while enabling automated error pattern extraction in subsequent stages. As a result, it ensures more accurate alignment with the reference and produces a robust transcription for further processing.

[1]https://github.com/m-bain/whisperX
[2]https://huggingface.co/openai/whisper-large-v3

B. Stage II: Error pattern extraction

In Stage II, we focus on identifying discrepancies between the ASR output and the reference transcription to extract what we define as "error patterns." To achieve this, we compare the ASR output and the reference transcription at the word level using several techniques. First, both the hypothesis (ASR output) and the reference are tokenized using the `jieba`[3] segmentation tool. After tokenization, we employ an algorithm to identify the non-matching portions by comparing the ASR output with the reference. Specifically, we detect the longest matching substrings between the two sequences. As illustrated in Fig. 1, the non-matching segments, representing the error patterns, are highlighted with markers "[src]" and "[tgt]" to distinguish them. This method ensures that the surrounding matching context is preserved while isolating the divergent sections. Once the error patterns have been identified, they are stored in an error database. This database enables the generation of additional contextual sentences based on these patterns, ultimately enhancing the model's generalization and robustness when dealing with similar errors in the future.

C. Stage III: Generate contextual corpus

In our task, the term "context" refers to the information surrounding a source sentence and is related to a specific piece of text. Context can be categorized based on its origin, encompassing the current sentence as well as the preceding and following k sentences. Both rule-based substitution and model-based generation methods are typically used to produce synthetic data, requiring a large volume of data to cover a wide range of errors. However, by leveraging the modeling capabilities of LLMs, we can now generate rich contextual data specifically tailored to high-quality ASR errors. Compared with traditional synthesis methods, our approach ensures that the error distribution aligns with the original dataset through rule-based error patterns, while model-based generation introduces diversity in sample contexts. To achieve this, we utilize two powerful LLMs: `Qwen2-72B-Instruct`[4] and `GPT-4o`[5]. The prompt used during this phase is as follows:

[3]https://github.com/fxsjy/jieba
[4]https://huggingface.co/Qwen/Qwen2-72B-Instruct
[5]https://openai.com/chatgpt/

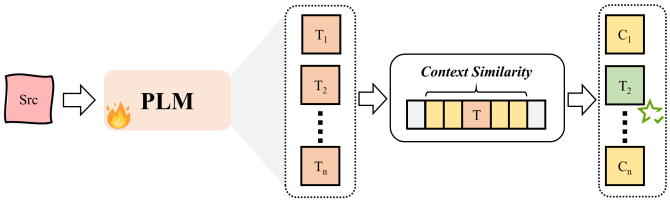

Fig. 2. Contextual similarity scoring for output refinement.

Prompt template:

You are given two versions of a text, marked as `[src]` and `[tgt]`. First, generate a full, correct contextual sentence based on the text labeled `[tgt]`. Then, replace the `[tgt]` part of the generated sentence with the content from `[src]`, keeping the rest of the sentence unchanged.

`[src:...]`
`[tgt:...]`
Please provide both sentences.

LLMs utilizes its extensive knowledge base to generate various contextual changes. By doing so, this method enhances the system's ability to adapt more effectively to various error patterns. Here, hyperparameters can be personalized to enable LLMs to generate multiple contexts.

D. Candidate selection

Once the training data is prepared, with the ASR output aligned to the reference transcription and errors identified, we use this data to fine-tune a PLM. The model's goal is to generate corrected transcriptions based on the input ASR text and its surrounding context. At the inference stage, we introduce a novel technique that enhances the adaptability of the generated content to the target's context. Specifically, we leverage the `sentence-transformer`[6] to select the output that is most semantically aligned with the given context. This approach is fully unsupervised and outperforms existing methods such as re-rank and beam search. For each candidate output, we compute the follow score:

Contextual Similarity: We compute a similarity score between the candidate and the context window.

$$\text{Sim}(S_T, C_{\text{context}}) = \frac{\mathbf{E}(S_T) \cdot \mathbf{E}(C_{\text{context}})}{\|\mathbf{E}(S_T)\| \|\mathbf{E}(C_{\text{context}})\|} \quad (2)$$

where $\mathbf{E}(S_T)$ and $\mathbf{E}(C_{\text{context}})$ represent the embeddings of the candidate S_T and the context window C_{context}, respectively. Context window is a hyperparameter that can be personalized according to different tasks.

III. Experiments

A. Datasets

To evaluate the effectiveness of our proposed model, we utilize two datasets generated from different ASR engines. For the single-hypothesis dataset, we focus solely on error pattern extraction to demonstrate the few-shot learning capability of our model.

- **LEMON** [22] is a multi-domain benchmark dataset for Chinese spelling correction (CSC). It is specifically designed to evaluate the generalization capabilities of CSC models across various open domains, providing a comprehensive assessment of their performance in diverse linguistic contexts.
- **Wang271K** [23] is constructed by introducing errors based on two criteria: visual similarity and phonetic similarity. The dataset serves as a robust foundation for training models to address common types of ASR errors encountered in real-world applications.
- **Classroom dialogue (CD)** is our collected dialogue dataset, we randomly sampled complete dialogue segments from the transcribed classroom dialogues, ensuring that the context remains intact. In total, we extracted 2,000 utterances, which were then annotated as demonstrated in Stage 1.

B. Baseline

To comprehensively evaluate the effectiveness of our proposed method, we compare it against two widely used baselines, representing different approaches for ASR error correction, including Seq2Seq models and prompt-based methods for large language models.

- **Original**: Refers to the WER of the original datasets when no error correction applies.
- **MacBERT [24]**: Fine-tuned on SIGHAN and Wang271K, MacBERT is an improved version of BERT with a novel MLM (Masked Language Modeling) correction pre-training task, designed to reduce the gap between pre-training and fine-tuning.
- **GPT-4o**: Following the prompt design outlined in [25], we adapted their approach by constructing prompts tailored for our task. Specifically, we input classroom dialogue directly into the ChatGPT API and compare the model's output with the ground truth to assess its performance.

C. Experimental settings

We conducted experiments using subsets of 50, 500, 2000 samples, as well as the full training set, from two widely recognized datasets: LEMON and Wang271k. The context window for candidate selection was set to 5. Training and evaluation were performed using an NVIDIA RTX 4090 GPU. The model's performance across different initial sample sizes is depicted in Fig. 3, highlighting the impact of varying the number of training examples on the final evaluation metrics. We conducted further evaluation using our own collected dataset, which was divided into 1800 samples for training and 200 samples for testing. This allowed us to assess the model's generalization ability to the data.

[6]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

TABLE II
COMPARISON OF RESULTS FOR THE LEMON, WANG271K AND CD DATASET WITH OTHER BASELINES.

Dataset	Model	Sample	WER (%)	Improvement (%)
LEMON	Origin	*Full*	4.63	-
	MacBERT	*Full*	12.82	-8.19
	GPT-4o	*Full*	3.72	+0.91
	EPIC	50	4.11	+0.52
		500	1.75	+2.88
		2000	1.02	+3.61
Wang271k	Origin	*Full*	12.40	-
	MacBERT	*Full*	0.38	+12.02
	GPT-4o	*Full*	9.26	+0.73
	EPIC	50	10.77	+1.63
		500	5.57	+6.83
		2000	0.96	+11.44
CD	Origin	1800	20.17	-
	BART	1800	8.63	+11.54
	MacBERT	1800	5.74	+14.43
	GPT-4o	1800	10.14	+10.03
	EPIC	50	12.51	+7.63
		500	7.53	+12.61
		1800	2.67	+17.5

TABLE III
EXPERIMENTAL RESULTS OF ABLATION ON DIFFERENT DATASETS.

Model Type	LEMON	Wang271k	CD	Average
Full Model	1.02	0.96	2.67	1.55
w/o Section II-B	1.75	1.24	2.85	1.95
w/o Section II-C	5.85	4.75	5.23	5.28
w/o Section II-D	1.30	1.15	2.90	1.78

D. Results analysis

The results in Table II clearly show the effectiveness of our model in correcting ASR errors across multiple datasets. Notably, our model achieved performance comparable to state-of-the-art (SOTA) models fine-tuned on full datasets, while consistently maintaining low WER values, even with limited training data. This suggests that the combination of error pattern extraction, contextual reconstruction, and candidate selection effectively captures the nuances of ASR outputs, making the model both stable and flexible in varying resource settings. Particularly on our custom CD dataset, where the training set was relatively small, our model demonstrated robustness by achieving near-identical results to models trained on significantly larger datasets. In comparison with LLMs like GPT-4o, although some improvements over the baseline were observed, GPT-4o often excessively modified sentences, which in some cases negatively impacted performance. Our model, in contrast, maintained an optimal balance between error correction and preserving the integrity of the original utterance. Moreover, MacBERT's strong performance on Wang271k underscored the benefits of domain-specific training. Despite the lack of extensive domain-specific fine-tuning, our model proved capable of achieving similar performance by effectively utilizing error patterns and contextual information.

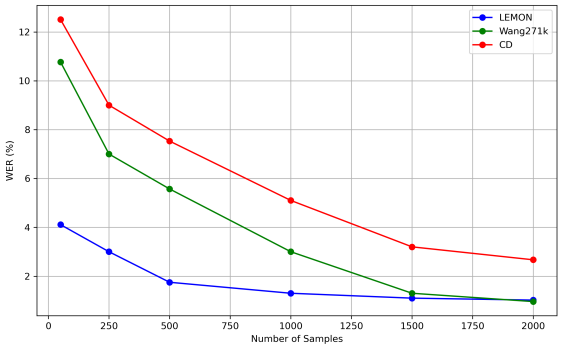

Fig. 3. The trend of WER under different training sample sizes.

E. Ablation studies

We conducted an ablation study by systematically removing key components and evaluating the resultant performance degradation, as shown in Table III. **w/o Section II-B** refers to removing the error extraction module. In this case, we retain sentence pairs that include both positive and negative samples, allowing the LLMs to automatically identify error types and generate additional similar examples. The absence of this module led to a notable decrease in performance, highlighting the importance of error pattern extraction for providing meaningful training samples. **w/o Section II-C** refers to removing the contextual dialogue generation process. Without this data augmentation step, the model was deprived of additional diverse examples that could help the model generalize better. This led to a clear drop in performance, illustrating that the absence of augmented contextual data significantly weakens the model's ability. **w/o Section II-D** tested the impact of context on correction performance. Inspired by [26], we evaluated this by shuffling the test set and removing the candidate selection module, treating the task as isolated sentence error correction. The results demonstrated that the inclusion of document-level context substantially enhanced the model's performance, while its absence led to significantly lower accuracy.

IV. CONCLUSIONS

This article proposes a new ASR error correction model **EPIC** for post-processing classroom dialogue transcription. The model utilizes LLMs to extract ASR error patterns, which can be effectively executed even with limited training data. At the same time, our model achieves results comparable to state-of-the-art methods trained on complete datasets. A large number of comparative experiments have demonstrated the robustness and adaptability of our method, making it a strong candidate for improving ASR correction in low resource scenarios. This method has been applied in practical classroom dialogue analysis and has a positive effect on downstream tasks.

REFERENCES

[1] Emily Jensen, Meghan Dale, Patrick J. Donnelly, Cathlyn Stone, Sean Kelly, Amanda Godley, and Sidney K. D'Mello, "Toward automated feedback on teacher discourse to enhance teacher learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2020, CHI '20, p. 1–13, Association for Computing Machinery.

[2] Yue Weng, Sai Sumanth Miryala, Chandra Khatri, Runze Wang, Huaixiu Zheng, Piero Molino, Mahdi Namazifar, Alexandros Papangelis, Hugh Williams, Franziska Bell, et al., "Joint contextual modeling for asr correction and language understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6349–6353.

[3] Meghan E. Dale, Amanda J. Godley, Sarah A. Capello, Patrick J. Donnelly, Sidney K. D'Mello, and Sean P. Kelly, "Toward the automated analysis of teacher talk in secondary ela classrooms," *Teaching and Teacher Education*, vol. 110, pp. 103584, 2022.

[4] A Suresh, J Jacobs, V Lai, C Tan, W Ward, JH Martin, and T Sumner, "Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application," *Association for the Advancement of Artificial Intelligence in Education*, 2021.

[5] Danner Schlotterbeck, Abelino Jiménez, Roberto Araya, Daniela Caballero, Pablo Uribe, and Johan Van der Molen Moris, ""teacher, can you say it again?" improving automatic speech recognition performance over classroom environments with limited data," in *Artificial Intelligence in Education*, Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova, Eds., Cham, 2022, pp. 269–280, Springer International Publishing.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 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)*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds., Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, Eds., Online, July 2020, pp. 7871–7880, Association for Computational Linguistics.

[8] Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiangyang Li, Edward Lin, and Tie-Yan Liu, "Fastcorrect: Fast error correction with edit alignment for automatic speech recognition," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, pp. 21708–21719, Curran Associates, Inc.

[9] Jinxi Guo, Tara N Sainath, and Ron J Weiss, "A spelling correction model for end-to-end speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5651–5655.

[10] Yichong Leng, Xu Tan, Rui Wang, Linchen Zhu, Jin Xu, Wenjie Liu, Linquan Liu, Xiang-Yang Li, Tao Qin, Edward Lin, and Tie-Yan Liu, "FastCorrect 2: Fast error correction on multiple candidates for automatic speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, Eds., Punta Cana, Dominican Republic, Nov. 2021, pp. 4328–4337, Association for Computational Linguistics.

[11] Rao Ma, Mark J. F. Gales, Kate M. Knill, and Mengjie Qian, "N-best t5: Robust asr error correction using multiple input hypotheses and constrained decoding space," in *INTERSPEECH 2023*. 2023, ISCA.

[12] Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér, "Whispering LLaMA: A cross-modal generative error correction framework for speech recognition," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali, Eds., Singapore, Dec. 2023, pp. 10007–10016, Association for Computational Linguistics.

[13] Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke, "Generative speech recognition error correction with large language models and task-activating prompting," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[14] CHEN CHEN, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng, "Hyporadise: An open baseline for generative speech recognition with large language models," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. 2023, vol. 36, pp. 31665–31688, Curran Associates, Inc.

[15] Rithik Sachdev, Zhong-Qiu Wang, and Chao-Han Huck Yang, "Evolutionary prompt design for llm-based post-asr error correction," 2024.

[16] Anisia Katinskaia and Roman Yangarber, "GPT-3.5 for grammatical error correction," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, Eds., Torino, Italia, May 2024, pp. 7831–7843, ELRA and ICCL.

[17] Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui, "Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction," 2023.

[18] Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang, "Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation," 2023.

[19] Takuma Udagawa, Masayuki Suzuki, Masayasu Muraoka, and Gakuto Kurata, "Robust asr error correction with conservative data filtering," 2024.

[20] Jingyuan Yang, Rongjun Li, and Wei Peng, "Asr error correction with constrained decoding on operation prediction.," in *INTERSPEECH*, Hanseok Ko and John H. L. Hansen, Eds. 2022, pp. 3874–3878, ISCA.

[21] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *INTERSPEECH 2023*, 2023.

[22] Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao, "Rethinking masked language modeling for Chinese spelling correction," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, Eds., Toronto, Canada, July 2023, pp. 10743–10756, Association for Computational Linguistics.

[23] Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang, "A hybrid approach to automatic corpus generation for Chinese spelling check," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, Eds., Brussels, Belgium, Oct.-Nov. 2018, pp. 2517–2527, Association for Computational Linguistics.

[24] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu, "Revisiting pre-trained models for chinese natural language processing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, Association for Computational Linguistics.

[25] Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang, "Connecting large language models with evolutionary algorithms yields powerful prompt optimizers," *arXiv preprint arXiv:2309.08532*, 2023.

[26] Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li, "Does multi-encoder help? a case study on context-aware neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, Eds., Online, July 2020, pp. 3512–3518, Association for Computational Linguistics.