Context Beyond Grammar: Synonym Substitution for Korean Grammatical Error Correction in Specialized Texts

Anonymous ACL submission

Abstract

Many previous studies on grammatical error correction (GEC) have primarily focused on language learner corpora, which consist of texts written by learners acquiring a non-native language. In this study, we address a GEC task that involves selecting contextually appropriate words in texts containing domain-specific vocabulary. We propose the `UniGEC` (Unified-Replacement GEC) dataset, which combines results from multiple models to determine the likelihood of substituting synonyms for specific keywords, based on token occurrence probabilities. Our experiments show that the `UniGEC` presents a more challenging task compared to language learner corpora. We observed that as the number of synonyms increases, the performance gap widens. Furthermore, we found significant performance variations across different domains, highlighting the need for further exploration of synonym substitution in specialized texts to expand the applicability of GEC tasks to a wider range of scenarios.

1 Introduction

Grammatical error correction (GEC) task involves identifying and correcting various types of errors in sentences (Bryant et al., 2023; Wang et al., 2021). With the advent of large language models (LLMs), there has been considerable effort to leverage their extensive pre-trained knowledge to enhance performance on this task (Davis et al., 2024; Zeng et al., 2024; Katinskaia and Yangarber, 2024; Zhang et al., 2023a). However, most existing GEC datasets in Korean have focused on relatively simple errors, such as those commonly made by language learners or typological mistakes frequently encountered online (Yoon et al., 2023; Koo et al., 2022; Lee et al., 2021; Min et al., 2020).

It can be effectively used in various tasks to improve sentence clarity (Bryant et al., 2023). In this study, we explored scenarios where GEC could be applied to specialized texts. Writing that conveys specific information contains complex vocabulary and poses challenges in understanding the context (Candlin and Plum, 2014; Ädel, 2010). Consequently, it is required to choose a more contextually appropriate word to ensure sentence clarity (Nirenburg and Nirenburg, 1988).

For instance, in the sentence "*The experiment proved the hypothesis under specific conditions*," there are no explicit grammatical errors, but depending on the context, the word *supported* might be more appropriate than *proved*[1]. Understanding synonyms is notably more challenging than identifying semantically unrelated words (Waring, 1997; Tinkham, 1993; Higa, 1963), and the task to distinguish them within texts containing advanced vocabulary remains unexplored. In this paper, we expanded the scope of GEC to include these cases where word selection depends on the context of the specialized texts. This approach informed the construction of `UniGEC` (Unified-Replacement GEC) dataset, with particular emphasis on leveraging multiple models to establish a robust foundation for the use of appropriate synonyms.

The process of constructing `UniGEC` is shown in Figure 1. We collected a corpus of research paper summaries from 8 domains, including *Social Science*, *Engineering*, and *Humanities*. First, we used LLMs to extract keywords from each text (Lee et al., 2023) to identify words that could be replaced with synonyms. To facilitate keyword extraction within each domain, we employed few-shot learning (Brown et al., 2020), providing LLMs with domain-specific samples during the instruction. By focusing on keywords consistently identified across multiple LLMs, we ensured more robust results and minimized the dependence on a single

[1] If the experiment provided evidence for the hypothesis under specific conditions, *supported* would be more appropriate, whereas *proved* would be used if the hypothesis was fully validated under all conditions.

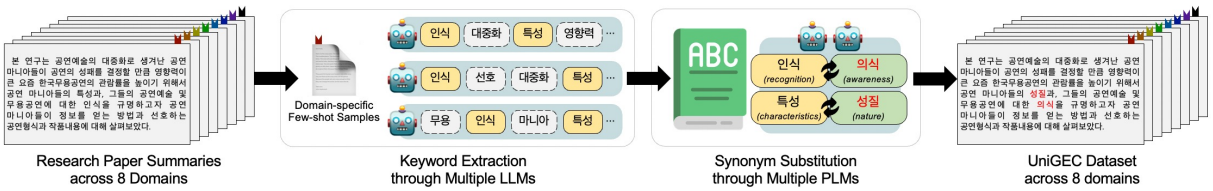

Figure 1: Process of constructing the proposed UniGEC dataset. It involves extracting keywords from research paper summaries across multiple LLMs, and replacing these keywords with synonyms or inserting typos.

model (Wang et al., 2024; Jiang et al., 2023).

The extracted keywords were replaced with other words from a predefined list of synonyms. In Figure 1, the word '인식(*recognition*)' was replaced with '의식(*awareness*).' While the word '의식' itself does not contain a grammatical error, its usage might feel awkward depending on the context of the specialized text. In the process of selecting which synonym to use from the list, we employed the concept of masked language modeling (Lewis et al., 2020; Devlin et al., 2019). By replacing the keywords in the original text with a [mask] token, we considered the occurrence probabilities of the surrounding tokens to determine the suitability of each synonym replacement (Chang et al., 2024; Li et al., 2023). We also leveraged probabilities from multiple pre-trained language models (PLMs) rather than relying on a single model, enabling us to achieve more stable results (Zou et al., 2024; Zhang et al., 2020).

We conducted correction task based on prompting techniques previously introduced and evaluated their performance with the UniGEC dataset. Unlike learner corpora (Yoon et al., 2023) commonly used in existing GEC tasks, we observed that the correction performance showed limited under the same prompting configurations. Through this, we emphasize the need to go beyond detecting explicit grammatical errors, highlighting the importance of identifying and revising contextually unnatural word choices in specialized texts. We will make UniGEC dataset publicly available to enable further research in this field[2].

2 Related Work

The GEC task has traditionally focused on language learner corpora, which arise during the process of non-native speakers learning a foreign language (Fang et al., 2023b; Takahashi et al., 2020; Bryant et al., 2019; Ng et al., 2014). For English, the primary focus of past research, these corpora have been drawn from essays written by learners or their online language use (Yannakoudakis et al., 2018; Dahlmeier et al., 2013). While GEC datasets have been developed for other languages like Chinese, German, and Russian, they also developed and utilized datasets derived from language learner corpora (Zhang et al., 2023b; Fang et al., 2023a; Katinskaia and Yangarber, 2023; Zhang et al., 2022).

For Korean, datasets have similarly been constructed by leveraging errors from language learning or by manually introducing noise (Lee et al., 2021; Min et al., 2020). Some studies employed human annotators to create realistic errors (Koo et al., 2022), while other categorized error types and conducted detailed analyses (Yoon et al., 2023). Many recent studies using LLMs have also been conducted on language learner corpora (Koo et al., 2024; Maeng et al., 2023).

In this work, unlike learner corpora that explicitly contain grammatical errors, we constructed a dataset focusing on how word choice depends on context. Recognizing that even native speakers may struggle with selecting the appropriate synonym, we aimed to ensure the correct use of synonyms by integrating the results from multiple models.

3 UniGEC: Dataset Construction

We used a research paper summary dataset divided into eight topics, ensuring the inclusion of domain-specific context. The detailed dataset preprocessing steps are provided in Appendix A.

3.1 Keyword Extraction

We conducted keyword extraction to determine which words in the given text should be replaced with synonyms. We employed few-shot learning (Brown et al., 2020), specifically providing domain information and domain-specific few-shot samples consisting of pairs of actual domain texts and their extracted keywords by human. This ap-

[2] https://anonymous.4open.science/r/UniGEC-895F/README.md

proach aimed to improve the capabilities of the model by incorporating additional information, rather than depending solely on its inherent parameters (Shi et al., 2024).

We use keywords that were consistently identified across multiple LLMs, rather than relying on the output of a single model. Let $K_{(i,j)}$ represent the set of keywords extracted from $\texttt{text}_i$ by the jth model, the mutual keywords K_i are as follows:

$$K_i = K_{(i,1)} \cap K_{(i,2)} \cap K_{(i,3)}, \quad (1)$$

The types of LLMs used, along with examples of the extracted keywords and their associated statistics, and details in domain-specific samples are provided in Appendix B.1.

3.2 Synonym Substitution

We retrieved synonym lists from the Naver Korean Dictionary[3], a popular resource among Korean speakers that offers definitions and synonyms for specific terms. For each keyword in the set $K_i = \{k_1, k_2, ..., k_n\}$, the corresponding synonym list S_i was obtained as follows:

$$\texttt{dict}(k_n) = \{k_n : s_{(n,1)}, s_{(n,2)}, ..., s_{(n,m)}\}, \quad (2)$$

$$S_i = \{\texttt{dict}(k_1), ..., \texttt{dict}(k_n)\}, \quad (3)$$

To decide which synonym to use for replacing each keyword, we adopted the concept of masked language modeling (Devlin et al., 2019). We replaced the keyword k_n in the $\texttt{text}_i$ with a `[mask]` token and then inserted one of the synonyms from $S_i[k_n]$. Let the synonym $s_{(n,m)}$ represent the selected one, they can be represented as follows:

$$s_{(n,m)} = \texttt{select}(S_i[k_n]), \quad (4)$$

$$\texttt{text}_i = \texttt{replace}(\texttt{[mask]}, s_{(n,m)}), \quad (5)$$

We then passed the $\texttt{text}_i$ through a PLM to obtain the token probability distribution $\texttt{probs}_i$. By excluding the dimension corresponding to the selected synonym, we calculated the product of the probabilities for surrounding tokens (Chang et al., 2024; Li et al., 2023), allowing us to consider the overall context when replacing the original keyword with its synonym.

In this case, we considered the probabilities from two PLMs to achieve more stable results. Let $\texttt{probs}_i \in \mathbb{R}^D$ is the averaged probabilities from both models, the *replaced probability* $p_{<i,k_n,s_{(n,m)}>}$ is as follows:

$$p_{<i,k_n,s_{(n,m)}>} = \prod_{d=1}^{D} \texttt{probs}_i^d, \quad \textit{if } d \textit{ is unrelated to the } s_{(n,m)}, \quad (6)$$

We calculated the $p_{<i,k_n,s_{(n,m)}>}$ for all synonyms corresponding to a given keyword. The final synonym replacement was determined by comparing the absolute differences between these probabilities and the *keyword probability* $p_{<i,k_n,k_n>}$[4]. If m' represents the index of the synonym with the largest difference, the keyword k_n is replaced with the synonym $s_{(n,m')}$.

$$m' = \texttt{argmax}(|p_{<i,k_n,s_{(n,m)}>} - p_{<i,k_n,k_n>}|), \quad \textit{for all } s_{(n,1)}, ..., s_{(n,m)} \textit{ in } \texttt{dict}(k_n), \quad (7)$$

We applied synonym substitution across the entire research paper summary texts. The details about the equations and process for synonym substitution are provided in Appendix B.2.

4 Experiment

4.1 Experimental Design

We conducted experiments using the `UniGEC` dataset to evaluate LLMs' ability to correct swapped synonyms in specialized texts. Following prior studies, we provided task instructions (Wu et al., 2023a) and applied zero-shot chain-of-thought (CoT) (Kojima et al., 2022) to encourage the models to leverage reasoning paths. Additionally, we employed task decomposition (Zhou et al., 2022), instructing the models to first identify unnatural words in the text before making corrections.

We selected four versions of LLMs for our experiments. `Qwen` (Qwen Team, 2024) and `Gemma` (Team et al., 2024) series are the recent models and perform well in Korean, even though they are multilingual. The details in the implementations and prompt configurations are provided in Appendix C.1 and D.

4.2 Main Results

The results of evaluating the correction of substituted synonyms across different domains are presented in Table 1. We set up to three keywords per

[3] https://ko.dict.naver.com

[4] The *keyword probability* is calculated in the same way as the *replaced probability*, as outlined in Equations (5)-(7), but with the synonym $s_{(n,m)}$ replaced by the keyword k_n. It represent the probability that a specific keyword is considered in the given context.

Method		Task Description		Zero-shot CoT		Task Decomposition	
Metric		BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
Learner Corpus	Gemma2-2b	69.34	61.77	63.14	54.74	64.63	57.82
UniGEC	Gemma2-2b	53.23	55.67	41.25	44.04	55.86	58.11
	Gemma2-9b	55.21	57.48	45.22	47.70	61.15	62.92
	Qwen2.5-1.5b	59.21	61.33	52.94	55.28	58.11	59.82
	Qwen2.5-7b	50.14	52.79	47.54	50.40	61.93	63.83

Table 1: Correction task performance with the prompting methods applied to each model. We compared UniGEC with the learner corpus (Yoon et al., 2023), presenting the results of the best model for the latter[5].

Model	Synonyms	Task Description		Zero-shot CoT		Task Decomposition	
		BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
Gemma2 -9b	2	57.53	59.54	46.45	48.88	64.41	65.98
	3	55.21	57.48	45.22	47.70	61.15	62.92
	4	53.96	56.30	44.45	46.98	59.56	61.53
Qwen2.5 -7b	2	52.75	55.15	50.78	53.29	65.76	67.39
	3	50.14	52.79	47.54	50.40	61.93	63.83
	4	48.69	51.45	46.91	49.74	59.78	61.88

Table 2: Performance differences in the correction task by varying the maximum number of synonyms replaced per text, focusing on the larger models.

text were replaced with synonyms. While smaller models performed reasonably well with only the task description, but larger models achieved better results as prompts were refined through task decomposition. In particular, Qwen2.5-7b achieved a GLEU score of 63.83, showing an improvement of 11.04 points over the simpler prompt.

5 Discussion

We conducted further experiments and analyses to explore various aspects of the UniGEC dataset constructed through synonym replacements.

Comparison based on the Nature of the Corpora We observed that the UniGEC performance was significantly lower than that of the learner corpus. This suggests that distinguishing contextually appropriate synonyms in specialized texts is more challenging than addressing the simpler vocabulary and error types found in language learner corpora. Therefore, it is essential to develop GEC techniques tailored to the unique characteristics of each corpus. We emphasize the need to expand synonym replacement tasks across a broader range of specialized texts to address this challenge.

Impact of Synonym Replacement Counts While Table 1 reported results based on replacing up to three keywords per text, we also experimented

[5]The results from the experiments with other models are provided in Appendix C.2.

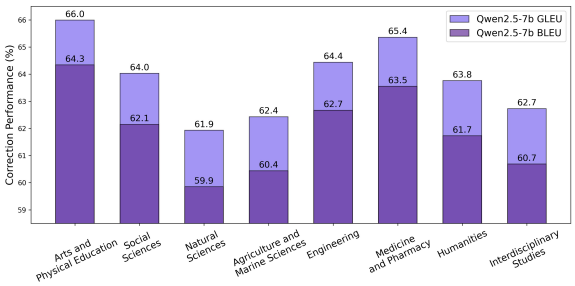

Figure 2: Performance differences across 8 domains in research paper summaries, focusing on the best-performed model Qwen2.5-7b in our experiments.

with varying the range to two and four keywords, as shown in Table 2. The results revealed that the correction performance consistently declined as the number of replaced keywords increased, with a maximum drop of 5.98 points. This highlights the difficulty models face in restoring the original text when more synonyms are replaced in that text.

Performance Variations across Domains When up to three keywords per text were replaced with synonyms, the results for each domain are presented in Figure 2. We observed significant variations depending on the domain-specific context. *Arts and Physical Education* achieved the highest scores, while *Natural Sciences* and *Agriculture and Marine Sciences* recorded the lowest, suggesting that synonym replacement is particularly challenging for scientific texts due to their specialized nature. The varied performance across domains further underscores the need for a more detailed analysis of synonym usage within each domain.

6 Conclusion

We introduce the UniGEC dataset, which performs synonym substitution by leveraging results from multiple models. This approach assumes that the GEC task, commonly applied to language learner corpora, can also be extended to specialized texts. The process involves extracting keywords and determining the probability of substituting synonyms. In our experiments, the results revealed that UniGEC is more challenging than language learner corpora. We observed that performance is influenced by the number of synonyms that can be substituted per text. Additionally, the performance variations across domains highlight the need for further research into synonym substitution for specialized texts, in order to expand GEC tasks to a broader range of scenarios.

Limitations

Nature of the Source Dataset Since our experiments were conducted using an exist paper summary dataset, we assumed that the texts were intentionally chosen to align with human intent, and treated them as the ground truth. This means that selecting a source dataset directly influences the approach to synonym substitution and significantly impacts results.

Absence of Direct Method We constructed a dataset based on synonym substitution, distinct from typical language learner corpora, but this study does not propose methods specifically designed for it. We plan to explore broader GEC scenarios using synonym substitution in diverse contexts and to propose methods tailored to these scenarios as future work.

Scalability of the Research While this study focuses on a Korean, expanding it to include English and other languages is essential for broader exploration. The careful selection of source datasets will also be crucial for other languages, and we believe our research on synonym substitution will offer valuable insights in this context.

Ethics Statement

We used multiple LLMs and PLMs in our approach, which may have influenced both the dataset construction and experimental results due to model biases. To mitigate this, we integrated results from various models during dataset creation to minimize such biases (Wang et al., 2024; Zou et al., 2024). Our goal was to develop a dataset that is not overly dependent on the results of any single model.

References

Annelie Ädel. 2010. Using corpora to teach academic writing: Challenges for the direct approach. *Corpus-based approaches to English language teaching*, pages 39–55.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Christopher N Candlin and Guenter A Plum. 2014. Engaging with the challenges of interdiscursivity in academic writing: researchers, students and tutors. In *Writing: Texts, processes and practices*, pages 193–217. Routledge.

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, and Tagyoung Chung. 2024. Explaining and improving contrastive decoding by extrapolating the probabilities of a huge and hypothetical LM. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8503–8526, Miami, Florida, USA. Association for Computational Linguistics.

Xinran Chen, Xuanang Chen, Ben He, Tengfei Wen, and Le Sun. 2024. Analyze, generate and refine: Query expansion with LLMs for zero-shot open-domain QA. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11908–11922, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Christopher Davis, Andrew Caines, Øistein E. Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of English learner text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11952–11967, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Fang, Jinpeng Hu, Derek F. Wong, Xiang Wan, Lidia S. Chao, and Tsung-Hui Chang. 2023a. Improving grammatical error correction with multimodal feature integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9328–9344, Toronto, Canada. Association for Computational Linguistics.

Tao Fang, Xuebo Liu, Derek F. Wong, Runzhe Zhan, Liang Ding, Lidia S. Chao, Dacheng Tao, and Min Zhang. 2023b. TransGEC: Improving grammatical error correction with translationese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3614–3633, Toronto, Canada. Association for Computational Linguistics.

Masanori Higa. 1963. Interference effects of intralist word relationships in verbal learning. *Journal of Verbal Learning and Verbal Behavior*, 2:170–175.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Anisia Katinskaia and Roman Yangarber. 2023. Grammatical error correction for sentence-level assessment in language learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 488–502, Toronto, Canada. Association for Computational Linguistics.

Anisia Katinskaia and Roman Yangarber. 2024. 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)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.

Lisa Kluge and Maximilian Kähler. 2024. Few-shot prompting for subject indexing of German medical book titles. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 141–148, Vienna, Austria. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Seonmin Koo, Jinsung Kim, Chanjun Park, and Heuiseok Lim. 2024. Search if you don't know! knowledge-augmented Korean grammatical error correction with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 96–125, Miami, Florida, USA. Association for Computational Linguistics.

Seonmin Koo, Chanjun Park, Jaehyung Seo, Seungjun Lee, Hyeonseok Moon, Jungseob Lee, and Heuiseok Lim. 2022. K-nct: Korean neural grammatical error correction gold-standard test set using novel error type classification criteria. *IEEE Access*, 10:118167–118175.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Myunghoon Lee, Hyeonho Shin, Dabin Lee, and Sung-Pil Choi. 2021. Korean grammatical error correction based on transformer with copying mechanisms and grammatical noise implantation methods. *Sensors*, 21(8).

Wanhae Lee, Minki Chun, Hyeonhak Jeong, and Hyunggu Jung. 2023. Toward keyword generation through large language models. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 37–40, New York, NY, USA. Association for Computing Machinery.

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.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Junghwan Maeng, Jinghang Gu, and Sun-A Kim. 2023. Effectiveness of ChatGPT in Korean grammatical error correction. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 464–472, Hong Kong, China. Association for Computational Linguistics.

Jinjong Min, Sungjun Jung, Sehee Jung, Sungmin Yang, Junsang Cho, and Sunghwan Kim. 2020. Grammatical error correction models for korean language via pre-trained denoising. *Quantitative Bio-Science*, 39(1):17–24.

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.

Sergei Nirenburg and Irene Nirenburg. 1988. A framework for lexical selection in natural language generation. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. GEE! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781, Mexico City, Mexico. Association for Computational Linguistics.

Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. 2020. Grammatical error correction using pseudo learner corpus considering learner's error tendency. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 27–32, Online. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Thomas Tinkham. 1993. The effect of semantic clustering on the learning of second language vocabulary. *System*, 21(3):371–380.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–51.

Robert Waring. 1997. The negative effects of learning words in semantic sets: A replication. *System*, 25(2):261–274.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023a. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023b. 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)*, pages 10743–10756, Toronto, Canada. Association for Computational Linguistics.

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.

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.

Min Zeng, Jiexin Kuang, Mengyang Qiu, Jayoung Song, and Jungyeul Park. 2024. Evaluating prompting strategies for grammatical error correction based on language proficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6426–6430, Torino, Italia. ELRA and ICCL.

Wangshu Zhang, Junhong Liu, Zujie Wen, Yafang Wang, and Gerard de Melo. 2020. Query distillation: BERT-based distillation for ensemble ranking. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 33–43, Online. International Committee on Computational Linguistics.

Yue Zhang, Leyang Cui, Enbo Zhao, Wei Bi, and Shuming Shi. 2023a. RobustGEC: Robust grammatical error correction against subtle context perturbation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16780–16793, Singapore. Association for Computational Linguistics.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. 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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9935–9951, Toronto, Canada. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Tianyuan Zou, Yang Liu, Peng Li, Jianqing Zhang, Jingjing Liu, and Ya-Qin Zhang. 2024. FuseGen: PLM fusion for data-generation based zero-shot learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2172–2190, Miami, Florida, USA. Association for Computational Linguistics.

A Dataset Descriptions

We collected a research paper summary dataset[6] and used only the valid dataset from the existing configuration. This dataset consists of abstracts summarized by human experts, ensuring that the text captures the overall content of the paper while incorporating domain-specific vocabulary. To ensure an appropriate length distribution, we removed the top 25% of texts that were either too short or too long. As a result, we used texts ranging from 154 to 239 lengths across all topics.

The topics were divided into eight categories: (1) *Arts and Physical Education*, (2) *Social Sciences*, (3) *Natural Sciences*, (4) *Agriculture and Marine Sciences*, (5) *Engineering*, (6) *Medicine and Pharmacy*, (7) *Humanities*, and (8) *Interdisciplinary Studies*. After preprocessing the texts, we standardized the dataset by selecting 140 texts from each topic, resulting in a total of 1,120 texts.

B Details in UniGEC Construction

B.1 Keyword Extraction Details

Selected LLMs We employed three instruction-tuned models trained on a Korean dataset[7][8], along with a recent multilingual model with string performance on Korean[9]. The temperature for keyword extraction was set to 0.2 (Chen et al., 2024).

Few-shot Samples We facilitated the keyword extraction for each LLM by incorporating domain-specific few-shot samples. Using the train split from the existing configuration of the source dataset we provided five sampled per text to enable few-shot learning. The prompt design for keyword extraction using few-shot samples was adapted from previous study (Kluge and Kähler, 2024).

Exceptional Cases To account for cases where no keywords were unanimously extracted by all models, we collected keywords extracted jointly by two models as $K_i^{'}$. Additionally, when increasing the maximum number of synonym substitutions as shown in Table 2, we supplemented the K_i with $K_i^{'}$ if the original one was insufficient.

$$K_i^{'} = (K_{(i,1)} \cap K_{(i,2)}) \cup (K_{(i,2)} \cap K_{(i,3)}) \cup (K_{(i,3)} \cap K_{(i,1)}) \setminus (3 * K_i), \quad (8)$$

B.2 Synonym Substitution Details

Selected PLMs We selected two types of models: one fine-tuned specifically for Korean[10] and another multilingual model with strong performance in Korean[11]. These models were used to derive token probabilities for input texts with substituted synonyms during inference stage.

Additional Explanations to Equations Using Equation (7), we described the process of replacing the keyword k_n with the synonym $s_{(n,m')}$. By applying this process to all keywords extracted from a given text, we obtained the list of substituted synonyms $S_i^{'}$. As previously defined, K_i and S_i represent the mutual keyword set and synonym list for $\texttt{text}_i$, respectively. This process are as follows:

$$S_{indices}^{'} = \{\text{Eq (7)}(k_n, \texttt{dict}(k_n)) \textit{ for } k_n \textit{ in } K_i \textit{ for } \texttt{dict}(k_n) \textit{ in } S_i\}, \quad (9)$$

$$S_i^{'} = \{\{k_n : s_{(n,m')}\} \textit{ for } n \textit{ in range}(len(K_i)) \textit{ for } m' \textit{ in } S_{indices}^{'}\}. \quad (10)$$

We prioritized keyword-synonym pairs from $S_i^{'}$ based on the largest absolute difference between the *replaced probability* and *keyword probability*, as defined in Equation (7). This approach aimed to avoid replacing all identified keywords with every possible synonym, instead selecting the synonym most contextually incongruous. In the experiments in Table 1, we used up to the top 3 keyword-synonym pairs, while in Table 2, we selected 2 to 4 pairs depending on the conditions.

Human Evaluation We conducted a human evaluation to determine whether the synonym sub-

[6] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=90

[7] https://huggingface.co/LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct

[8] https://huggingface.co/nlpai-lab/KULLM3

[9] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[10] https://huggingface.co/klue/bert-base

[11] https://huggingface.co/FacebookAI/xlm-roberta-base

Domains	Rate #1	Rate #2	Rate #3
Arts and Physical Education	2.4	2.0	2.2
Social Sciences	2.3	2.3	2.3
Natural Sciences	1.9	1.7	1.8
Agriculture and Marine Sciences	2.3	2.2	2.2
Engineering	2.2	2.1	1.8
Medicine and Pharmacy	2.2	2.1	1.8
Humanities	2.5	2.3	2.3
Interdisciplinary Studies	2.0	1.8	1.8

Table 3: Confusion scores in synonym-substituted texts for all domains, with higher scores indicating greater confusion. The average scores are reported.

stitution in the UniGEC dataset effectively create potential confusion with the original text as intended. Three native university graduates fluent in Korean volunteered for this evaluation. We asked them to rate whether the modified text could cause confusion if they were asked to write the original version. We provided each rater with 10 texts from each domain, for a total of 80 texts. The results of this evaluation are presented in Table 3.

Higher scores indicate that the synonym-replaced text is more confusing compared to the original, suggesting that the synonym replacement process effectively created texts that could confuse even human raters. The results showed that most domains had scores near or above 2 out of 3, reflecting a consistent level of difficulty. However, in the *Natural Sciences* domain, all raters gave lower scores, indicating that the complexity of synonym replacement may depend on the domain.

C Details in Experiments

C.1 Implementation Details

Prompt Configurations Unlike common GEC tasks that address explicit grammatical errors, correcting swapped synonyms based on context required modifications to the standard prompts. Furthermore, to mitigate the problem of over-correction (Wu et al., 2023b), where generative models tend to make unnecessary edits, we modified the prompt configurations accordingly.

Experimental Setup The temperature for the correction task was set to 0 (Song et al., 2024). We used BLEU and GLEU scores, which are commonly used metrics in GEC research (Koo et al., 2024; Yoon et al., 2023). To facilitate efficient inference with the LLMs used in our experiments, we utilized the vLLM library (Kwon et al., 2023).

Model		Task Description		Zero-shot CoT		Task Decomposition	
Metric		BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
Learner Corpus	Gemma2-2b	69.34	61.77	63.14	54.74	64.63	57.82
	Gemma2-9b	60.98	50.13	57.75	45.46	65.42	57.78
	Qwen2.5-1.5b	66.64	57.01	63.11	49.97	68.14	60.08
	Qwen2.5-7b	58.49	47.70	60.42	50.50	63.84	55.56

Table 4: Correction task performance of all models using the learner corpus.

C.2 Remaining Experimental Results

Results of Learner Corpus We present the full results of the four models using the learner corpus in Table 4. We selected the Kor-Learner dataset from the original one (Yoon et al., 2023), as it effectively represents the typical characteristics of language learner corpora. When comparing the results with those from UniGEC, we found that our dataset was more challenging in simpler prompt configurations, as outlined in the task description. However, when the task was broken down into its core components for inference, our dataset exhibited higher scores.

Performance Variations across Domains In addition to the results presented in Figure 2, we observed consistent performance differences for all models and metrics. The trends across all eight domains resembled those in Figure 2, with domains using more technical terminology, such as containing *Sciences*, showing notably lower performance.

D Prompt Templates

- Keyword extraction

> Your task is extracting the keywords from the given sentences.
>
> You will be provided with the text written on the topic of "{*domain_name*}".
> Please refer these examples, do not copy them for the generated results.
>
> # domain-specific few-shot samples
> sentences: {*sentences*}
> keywords: {*keywords*}
> ...
>
> Please extract the top 8 most significant keywords from the sentences below. Always answer in Korean without any explanations.
> sentences: {*input_text*}
> keywords:

- Task description for the leaner corpus

> Do grammatical error correction on all the following sentences. Always answer in Korean, without any explanations.
> input: {*input_text*}
> output:

- Task description for UniGEC

> Please revise any unnatural words in the given sentences to better fit the context, while keeping the rest unchanged as much as possible. Always answer in Korean, without any explanations.
> input: {*input_text*}
> output:

- Zero-shot CoT for the learner corpus

> # phase 1
> Do grammatical error correction on all the following sentences. Let's think step by step. Always answer in Korean.
> input: {*input_text*}
> reasoning path:
>
> # phase 2
> reasoning path: {*reasoning_path*}
> Do grammatical error correction that fit the given sentences. Let's think step by step. Always answer in Korean, without any explanations.
> input: {*input_text*}
> output:

- Zero-shot CoT for UniGEC

> # phase 1
> Please revise any unnatural words in the given sentences to better fit the context, while keeping the rest unchanged as much as possible.
> input: {*input_text*}
> reasoning path:
>
> # phase 2
> reasoning path: {*reasoning_path*}
> Please revise any unnatural words in the given sentences to better fit the context, while keeping the rest unchanged as much as possible. Let's think step by step. Always answer in Korean, without any explanations.
> input: {*input_text*}
> output:

- Task decomposition for the learner corpus

> # phase 1
> Please detect words with any grammatical errors in the given sentences. Always answer in Korean.
> input: {*input_text*}
> reasoning path:
>
> # phase 2
> reasoning path: {*reasoning_path*}
> Based on detected words, do grammatical error correction that fit the given sentences. Always answer in Korean, without any explanations.
> input: {*input_text*}
> output:

- Task decomposition for UniGEC

> # phase 1
> Please detect any unnatural words in the given sentences according to the context. Always answer in Korean.
> input: {*input_text*}
> reasoning path:
>
> # phase 2
> reasoning path: {*reasoning_path*}
> Based on detected words, please revise any unnatural words in the given sentences to better fit the context, while keeping the rest unchanged as much as possible. Always answer in Korean, without any explanations.
> input: {*input_text*}
> output: