Interspeech 2024
1-5 September 2024, Kos, Greece

Spoken-to-written text conversion with Large Language Model

HyunJung Choi [1*], *Muyeol Choi* [2*], *Yohan Lim* [1], *Minkyu Lee* [2], *Seonhui Kim*[1], *Seung Yun*[2], *Donghyun Kim*[2], *SangHun Kim*[2]

[1]University of Science and Technology, South Korea
[2]Electronics and Telecommunications Research Institute, South Korea
{Chocohj, mychoi, johnlim, mk, ksh05, syun, dawnkann, ksh}@etri.re.kr

Abstract

The improvement in end-to-end speech recognition systems has enhanced the readability of results, making it easier for users to understand texts and reducing translation errors. Korean uses both written and spoken forms, making it crucial to standardize pronunciation notation for high readability. Inverse Text Normalization (ITN) technology, which converts pronunciation into readable written form, can be applied in preprocessing training corpora or post-processing speech recognition outcomes. Recent Korean ITN research utilizes transformer models based on training data with both notations, facing performance degradation due to data scarcity. This paper proposes using Large Language Models for ITN to address this issue, overcoming the performance decline from limited data. The proposed method showed an 12.6% reduction in Error Reduction Rate (ERR).

Index Terms: Inverse Text Normalization, Spoken-to-Written, Large Language Model

1. Introduction

Highly readable Automatic Speech Recognition (ASR) outcomes can mitigate linguistic ambiguity, enabling users to comprehend texts with ease and clarity. When leveraging speech recognition results for automatic translation, the readability of the text plays a crucial role in reducing translation errors. Such readability-enhanced recognition results serve as indicators of the performance improvements in end-to-end ASR systems.

Korean employs both written (formal) and spoken (phonetic) forms. The written form is predominantly used in formal documents and represents the standard method for reading and writing. In contrast, the spoken form is a method that transcribes sounds as they are heard, without the use of symbols or abbreviations. The outcomes of end-to-end ASR systems are significantly influenced by the form of transcription used in training data, and it is crucial to consistently convert spoken forms to written forms in transcriptions to achieve readable recognition results. For instance, when an ASR system is trained on the numeral '3' in both its written form ('3') and its spoken form ('삼'), the recognition results might exhibit a mixture of these forms. While users can easily understand smaller numbers in their spoken form, readability decreases as the magnitude of the number increases, potentially causing inconvenience. For example, interpreting '2,579,347' as '이백오십칠만 구천삼백사십칠 (Two million five hundred seventy-nine thousand three hundred forty-seven)' requires [1]more effort from the user compared to its written form. A similar issue is observed in the transcription of loanwords in Korean. The spoken form of 'Personal Computer' is '피씨', whereas its written form uses the English acronym 'PC'. The phonetic component '씨' in '피씨' shares pronunciation with a Korean honorific, leading to ambiguity in contexts where it's unclear whether it refers to a computer or possibly a person named 'P'. Thus, there's a trend towards using written forms in transcriptions to enhance readability and reduce ambiguity.

Traditional ASR systems utilize transcriptions in spoken form. During the pre-processing phase, text normalization (TN) [1, 2, 3] processes are employed to construct corpora. As neural network training began to be applied to ASR systems, there was a shift towards using transcriptions in written form instead of spoken form. This necessitated ITN[4, 5, 6, 7, 8, 9] to overcome the discrepancies between pre-existing corpora in spoken form and the newly adopted written form. ITN can be used both in the preprocessing of mixed-form transcription corpora and in the post-processing of ASR results. Research applying ITN in the construction of speech corpora includes methods for generating both written and spoken forms by ASR systems[10] and converting from spoken to written form, considering the specific characteristics of a language[11]. Neural network-based ITN approaches are superior in learning the complex relationships between spoken and written forms with lower costs through a data-driven approach, compared to Finite State Transducer (FST)-based methods[12, 13, 14]. Research has been conducted to enable neural network-based ITN models to handle ASR-generated spoken form text that includes irregular insertions of interjections or errors, by using ASR output as training data[15]. This study explores methodologies to enhance the robustness and accuracy of ITN and reduce errors through data augmentation, semi-supervised learning, and post-alignment methods. A transformer model[16]-based ITN approach, focused on converting the spoken form of numbers and English into written form, was proposed and demonstrated to also improve the performance of automatic translation[17]. Reliance on training data that concurrently annotates spoken and written forms can limit model performance in cases of data scarcity. Additionally, encountering sentence structures different from learned patterns can pose challenges in the inference process, necessitating access to diverse and extensive training data. Data augmentation is often employed to address data scarcity, but it has limitations in extending the

* These authors contributed equally to this work.

10.21437/Interspeech.2024-376

(a) Dual transcription structure : (written form)/(spoken form)
(b) Dual transcription example : 주인공은 (35)/(삼십오)세 설계사 (*The protagonist is a (35)/(thirty-five)-year-old designer.*)
(c) Data format for Finetuning
: **<SPOKEN>**[BOS]이천이십사년대 지디피 성장률 이퍼센트를 예상했다.[EOS]**<WRITTEN>**[BOS] 2024년 GDP 성장률 2%를 예상했다.[EOS]
*(**<SPOKEN>**[BOS] It was anticipated that the **Jee-Dee-Pee** growth rate would be **two percent** in the **two thousand twenty-fourth** year.[EOS]*
*__<WRITTEN>__[BOS] It was anticipated that the **GDP** growth rate would be **2%** in **2024**.[EOS])*

Figure 1. *Examples of dual transcription and data format*
(The part written in italics is the English translation of the Korean sentence.)

characteristics of existing data, and the accumulation of errors during the augmentation process can increase the likelihood of system errors. Therefore, for the effective development of ITN systems, new methodologies are needed to ensure data diversity and minimize error accumulation.

In this paper, we propose an ITN approach based on LLMs to address these challenges. LLMs are language models trained on extensive datasets, capable of performing a wide array of natural language processing tasks. Having learned the intricate patterns of language from vast amounts of textual data, LLMs excel in various NLP tasks, including natural language understanding (NLU) and generation (NLG). They are also widely employed in applications such as machine translation, summarization, and question-answering systems. Prominent foundational models of LLM include OpenAI's GPT series (GPT-1[18], GPT-2[19], GPT-3[20], GPT-4[21]), Meta's Llama (Llama1[22], Llama2[23]), and Google's PaLM[24].

Recent research trends indicate that studies are being conducted on error correction techniques that improve the accuracy of text in the post-processing phase of ASR by utilizing LLMs[25, 26, 27]. Based on the performance improvements observed with LLMs fine-tuned on various NLP tasks, this study proposes generating a Korean Large Language Model and then applying this Korean LLM to ITN. This approach demonstrates that leveraging the knowledge acquired from large datasets can overcome the challenges associated with limited annotated training data. The proposed LLM-based ITN method enables effective responses to new forms of input not present in the training data, excels at understanding complex contexts, and reduces ambiguity, thus resolving potential issues in ITN performance.

2. Proposed Method

1.1. Korean Large Language Model

In this study, we propose the development of a Korean Large Language Model and its application to ITN. The Korean LLM model employed in this research was trained from scratch using a new Korean corpus, referencing the structure of GPT-2 small[19]. The GPT-2 model is a transformer-based model pre-trained on a large corpus of text data through self-supervised learning, demonstrating exceptional capability in predicting the next token in a sequence of text.

The previously developed Korean LLM, KoGPT[28], which has around 6 billion parameters, poses significant fine-tuning challenges in environments with limited training resources. To overcome these constraints, this study has developed a Korean LLM based on the GPT-2 small architecture, enabling efficient use of resources.

1.2. Korean Inverse Text Normalization Model (K-ITN)

The K-ITN model, designed for ITN tasks, has been fine-tuned on the foundation of a Korean Large Language Model. This fine-tuning specifically aims to learn the necessary patterns for converting spoken forms to written forms. The process of converting spoken forms to written forms by the K-ITN model is illustrated in Figure 2. The K-ITN model consists solely of decoders. It accepts text sequences as input, which are then tokenized into words or subwords. Each token is transformed into a fixed-size vector through an embedding layer, and positional encoding is added before the token is fed into the decoder blocks. The output of the model is the probability distribution of the next token for a given input sequence, and the text sequence is generated incrementally by selecting the token with the highest probability from the output distribution.

The dataset used for fine-tuning the model is structured in a dual transcription form, where spoken and written forms are paralleled, as illustrated in Figure 1(a). In cases where the spoken and written forms differ within a sentence, they adopt a parallel transcription form as shown in Figure 1(b). The input to the model is a text sequence structured as "<SPOKEN>[BOS] Spoken form sentence [EOS]<WRITTEN>[BOS] Written form sentence [EOS]," as depicted in Figure 1(c). <SPOKEN> and <WRITTEN> are tokens indicating the start of the spoken and written form, respectively. [BOS] signifies the beginning of a sentence, and [EOS] denotes the end of a sentence. When the set of Spoken token(S) is $\{S_1, S_2, \dots, S_N\}$ and the set of Written token(W) is $\{W_1 W_2, \dots, W_M\}$, the loss calculation formula is as follows.

$$\mathcal{L} = -\sum_{i=1}^{M} \log(\mathbb{P}(w_i | S, W_{<i}, \theta))$$

For the fine-tuning process of the K-ITN model, it becomes crucial to focus solely on the sentences in the written form during the backward propagation phase. Therefore, to facilitate the model's accurate acquisition of the written form, the backward propagation step excludes the spoken tokens present within the sentence.

3. Experiment & Results

The corpus employed for the Korean LLM incorporates an extensive array of data sources furnished by AIHub[29], featuring a comprehensive corpus composed of texts from purchased books and specialized texts pertinent to the fields of medicine and law. Additionally, it integrates parliamentary transcripts and newspaper articles, alongside content from

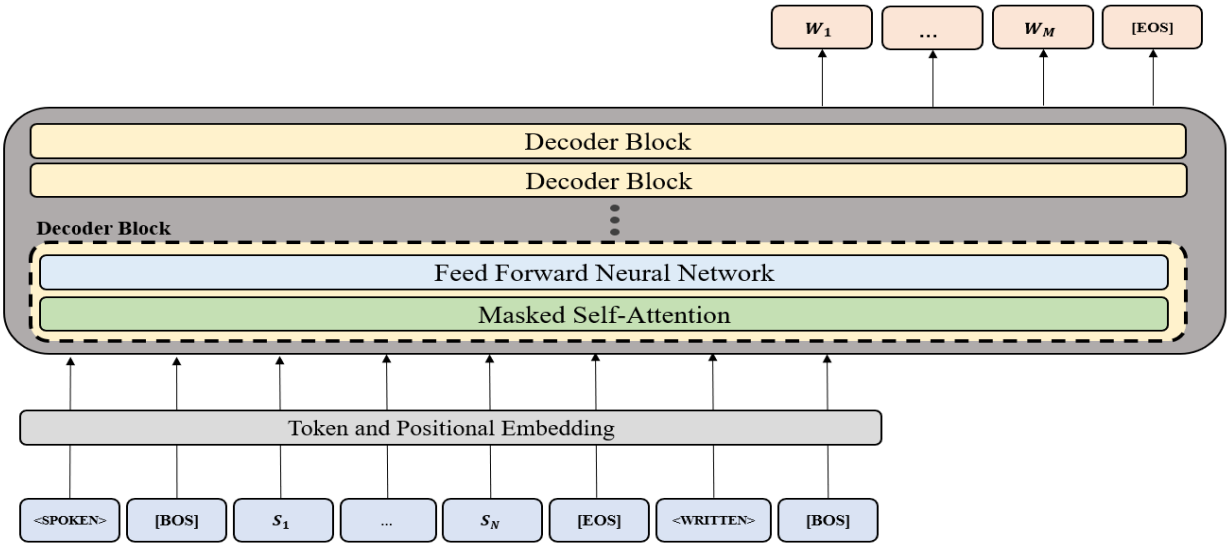

Figure 2. *Architecture of Proposed method K-ITN*

Korean Wikipedia and Namuwiki, all collected from Korpora [30]. The model was trained using a dataset comprising 7.6 billion tokens, with a total of 116 million parameters.

The fine-tuning utilized a dataset comprising 4.3 million sentences from AIHub Korean Lecture (AIHub-KL) speech dataset[31] and ETRI Korean Common (ETRI-KC) speech dataset[32]. For performance evaluation, a dataset of 3,000 news articles not included in the training data was utilized. The performance of the K-ITN model was compared against a baseline ITN model based on transformer[16]. The evaluation distinguished between Target characters, which should be transformed by ITN, and Non-target characters, which should not be transformed. In this experiment, Korean syllables were defined as the basic characters. The Inversion Character Error Rate (I-CER) measured the error rate of Target syllables, while the Non-Inversion Character Error Rate (NI-CER) measured the error rate of Non-target syllables. I-CER was used as a metric to assess how accurately the model recognized and transformed Targets in the input test, and NI-CER was used as a metric to evaluate how well the model preserved parts that should not be transformed during the conversion process. The performance of the baseline and the proposed method are presented in Table 1.

Table 1. *Comparison between baseline and proposed method*

	I-CER(%)	NI-CER(%)
Baseline (No LLM)	9.02	8.39
K-ITN (Small)	8.46	7.49
K-ITN (Large)	7.88	7.24

In the comparative experiment, an investigation was conducted alongside to explore the impact of the amount of tuning data used during the fine-tuning process of Korean LLMs on model performance. The fine-tuning dataset, comprised of a total of 4.3 million sentences, was used in two different proportions: the K-ITN(Small) model was trained using only 10% of the dataset, and the K-ITN(Large) model was trained using the entire 100% of the dataset. The purpose of this was to assess the feasibility of applying the model to other languages with relatively small amounts of parallel corpora data. The I-CER and NI-CER for the baseline model, which is based on transformers, were 9.02% and 8.39%, respectively. In comparison, the best performance of K-ITN achieved I-CER and NI-CER of 7.88% and 7.24%, respectively, with an Error Reduction Rate (ERR) of 12.6% and 13.7%. This demonstrates the effectiveness of the LLM-based ITN over traditional methods. Examining the results of the fine-tuning model experiments according to the amount of data, the K-ITN(Small), which used only 10% the fine-tuning data, achieved I-CER and NI-CER of 8.46% and 7.49%, respectively, with an ERR of 6.2% and 10.7%. Even with only 10% of the data, the performance was better than the baseline model, highlighting the robust performance of the proposed K-ITN model. This highlights the potential for applying multilingual ITN models for under-resourced languages with relatively small amounts of tuning data.

4. Qualitative Comparisons

The examples in Table 2 illustrate a qualitative comparison between the outcomes of transformer-based ITN models and LLM-based ITN models. Notably, the results from the K-ITN model demonstrate its ability to resolve ambiguities arising in spoken forms. In the first example, the spoken form notation ‘제이군(Jay-gun)’ refers to 'a male named J'. The Korean transliteration of the alphabet 'J' is '제이', where the Korean character '이' sounds identical to the pronunciation of the number '2'. The transformer-based ITN model incorrectly converts the phonetically similar '이' to the number '2', leading to the unintended output 'The Second Group'. In contrast, leveraging the capabilities of LLM, the K-ITN successfully interprets the meaning and accurately converts it to 'a male named J', showcasing its ability to discern and preserve the intended semantic content.

In the second example, the spoken form '십이개국 (*Ship-ee-gae-guk*, meaning: twelve countries)' was correctly interpreted as '십이', translating to the number '12'. The spoken form

Table 2. Examples of ITN results for the Baseline and K-ITN
(The part written in italics is the English translation of the Korean sentence.)

Input	증상은 대개 중년 이후에 시작되지만 **제이군처럼** 어린 나이에 나타나기도 합니다. *(Symptoms usually begin in middle age, but like **Jay-gun(meaning: a male named J),** they can also appear at a young age.)*
Baseline (No LLM)	증상은 대개 중년 이후에 시작되지만 **제 2 군처럼** 어린 나이에 나타나기도 합니다. *(Symptoms usually begin in middle age, but **like the second group**, they can also appear at a young age.)*
Proposed method (K-ITN Large)	증상은 대개 중년 이후에 시작되지만 **J 군처럼** 어린 나이에 나타나기도 합니다. *(Symptoms usually begin in middle age, but **like J**, they can also appear at a young age.)*
Input	**이유** 회원국 중 유럽의 **십이개국**에 대해 분석을 시행할 예정이다. *(Plans are in place to conduct an analysis on **twelve countries** in Europe among **Yi-yu(meaning: The European Union)** member states.)*
Baseline (No LLM)	**이유** 회원국 중 유럽의 **12 개국**에 대해 분석을 시행할 예정이다. *(Plans are in place to conduct an analysis on **12 countries** in Europe among **Yi-yu** member states.)*
Proposed method (K-ITN Large)	**EU** 회원국 중 유럽의 **12 개국**에 대해 분석을 시행할 예정이다. *(Plans are in place to conduct an analysis on **12 countries** in Europe among **EU** member states.)*
Input	경기도 **이천에서** 도자기 축제가 열렸다. *(A pottery festival was held in **Icheon**, Gyeonggi Province.)*
Baseline (No LLM)	경기도 **2000 에서** 도자기 축제가 열렸다. *(A pottery festival was held in Gyeonggi **2000**.)*
Proposed method (K-ITN Large)	경기도 **이천에서** 도자기 축제가 열렸다. *(A pottery festival was held in **Icheon**, Gyeonggi Province.)*

'이유(*Yi-yu*)', given the context with terms like 'Europe' or 'member countries', should be converted to 'EU', symbolizing the European Union.

However, the transformer-based model interpreted '이유', which has the same pronunciation as the Korean word for 'reason', in its literal meaning and failed to convert the spoken form to the written form 'EU'. On the other hand, the K-ITN model, considering the context, successfully converted the spoken form into the written form 'EU'.

The final example effectively illustrates the reason for the improved performance of the K-ITN model in handling ambiguous words with identical spoken forms. Within a sentence, "이천" has the same spoken form as the number "2000" and also matches the written form of a place name in Korea. While the transformer-based ITN model simply converted "이천" to the number "2000", the K-ITN model, utilizing the contextual clue "경기도", identified "이천" as a proper noun indicating a place name. Consequently, it preserved the original spoken form of "이천" without converting it to a number. This demonstrates that the K-ITN model, based on a LLM, surpasses traditional transformer-based ITN models in resolving ambiguities arising from spoken forms by considering the surrounding context to accurately convert spoken forms into their correct written forms.

5. Conclusion

In this study, the proposed K-ITN model has achieved significant results in the task of Korean ITN, in conjunction with LLMs. The K-ITN model, fine-tuned using Korean LLMs, has successfully addressed the ambiguity in spoken form notation and accurately converted to the correct written form by considering the context, compared to the existing transformer-based ITN models. Notably, even when the spoken forms are identical, the model leverages contextual clues to guide the correct conversion, and by accurately processing complex numerical expressions, it has shown improvements in the ERR by 12.6% and 13.7%, respectively, over conventional transformer-based models in performance evaluation.

Furthermore, even when the amount of fine-tuning data was limited to 10%, it demonstrated superior performance compared to transformer-based ITN models, presenting the potential for developing ITN models for sparse languages with relatively little data. These results suggest that ITN using LLMs can operate effectively in any linguistic environment, compared to traditional methods. Moreover, by resolving ambiguity and enhancing context understanding capabilities, the reliability and accuracy in the conversion process from spoken to written forms are significantly improved, setting a new standard for inverse text normalization tasks in various languages.

6. Acknowledgements

This study was supported by an Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean Government (24ZB1100, Core Technology Research for Self-improving Integrated Artificial Intelligence Systems).

7. References

[1] Allen, Jonathan, et al. From text to speech: The MITalk system. Cambridge University Press, 1987.

[2] Sproat, Richard, and Navdeep Jaitly. "RNN approaches to text normalization: A challenge." arXiv preprint arXiv:1611.00068 (2016).
[3] Pramanik, Subhojeet, and Aman Hussain. "Text normalization using memory augmented neural networks." Speech Communication 109 (2019): 15-23.
[4] Sunkara, Monica, et al. "Neural inverse text normalization." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.
[5] Paul, Debjyoti, et al. "Improving Data Driven Inverse Text Normalization using Data Augmentation and Machine Translation." INTERSPEECH. 2022.
[6] Pandey, Laxmi, et al. "Improving data driven inverse text normalization using data augmentation." arXiv preprint arXiv:2207.09674 (2022).
[7] Gaur, Yashesh, et al. "Streaming, fast and accurate on-device inverse text normalization for automatic speech recognition." 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023.
[8] Antonova, Alexandra, Evelina Bakhturina, and Boris Ginsburg. "Thutmose tagger: Single-pass neural model for inverse text normalization." arXiv preprint arXiv:2208.00064 (2022).
[9] Tan, Sharman, et al. "Four-in-One: a joint approach to inverse text normalization, punctuation, capitalization, and disfluency for automatic speech recognition." 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023.
[10] Ihori, Mana, et al. "Transcribing speech as spoken and written dual text using an autoregressive model." Proc. INTERSPEECH 2023 (2023): 461-465.
[11] Ihori, Mana, Akihiko Takashima, and Ryo Masumura. "Parallel corpus for Japanese spoken-to-written style conversion." Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.
[12] Shugrina, Maria. "Formatting time-aligned ASR transcripts for readability." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010.
[13] Zhang, Yang, et al. "Nemo inverse text normalization: From development to production." arXiv preprint arXiv:2104.05055 (2021).
[14] Ebden, Peter, and Richard Sproat. "The Kestrel TTS text normalization system." Natural Language Engineering 21.3 (2015): 333-353.
[15] Kim, Juntae, Minkyu Lim, and Seokjin Hong. "Improving Robustness of Neural Inverse Text Normalization via Data-Augmentation, Semi-Supervised Learning, and Post-Aligning Method." arXiv preprint arXiv:2309.08626 (2023).
[16] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[17] Choi, HyunJung, et al. "Spoken-to-written text conversion for enhancement of Korean-English readability and machine translation." ETRI Journal 46.1 (2024): 127-136.
[18] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).
[19] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
[20] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
[21] Achiam, Josh, et al. "GPT-4 technical report." arXiv preprint arXiv:2303.08774 (2023).
[22] Touvron, Hugo, et al. "LLaMA: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).
[23] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
[24] Chowdhery, Aakanksha, et al. "PaLM: Scaling language modeling with pathways." Journal of Machine Learning Research 24.240 (2023): 1-113.
[25] Radhakrishnan, Srijith, et al. "Whispering LLaMA: A cross-modal generative error correction framework for speech recognition." arXiv preprint arXiv:2310.06434 (2023).
[26] Chen, Chen, et al. "HyPoradise: An open baseline for generative speech recognition with large language models." Advances in Neural Information Processing Systems 36 (2024).
[27] Li, Yuang, et al. "Using Large Language Model for End-to-End Chinese ASR and NER." arXiv preprint arXiv:2401.11382 (2024).
[28] Kim, Ildoo, et al. "KoGPT: Kakaobrain korean (hangul) generative pre-trained transformer." Opgehaal van https://github.com/kakaobrain/kogpt (2021).
[29] AIHub, https://www.aihub.or.kr/
[30] Korpora, https://kli.korean.go.kr/corpus/main/requestMain.do
[31] AIHub, Aihub Korean lecture speech dataset, 2020. Last accessed on March 11, 2024.
[32] ETRI, Etri Korean common speech dataset, 2004. Last accessed on March 11, 2024.