Received 13 July 2023, accepted 25 August 2023, date of publication 30 August 2023, date of current version 8 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3310257

RESEARCH ARTICLE

Uncovering the Risks and Drawbacks Associated With the Use of Synthetic Data for Grammatical Error Correction

SEONMIN KOO[1], CHANJUN PARK[1,2], SEOLHWA LEE[3], JAEHYUNG SEO[1], SUGYEONG EO[1], HYEONSEOK MOON[1], AND HEUISEOK LIM[1]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea
[2]Upstage, Yongin-si, Gyeonggi-do 17006, South Korea
[3]Department of Computer Science, Technical University of Darmstadt, 64289 Darmstadt, Germany

Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2018-0-01405) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and this work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques), and This work was supported by the Technology development Program (1425173998) funded by the Ministry of SMEs and Startups (MSS, Korea).

ABSTRACT In a Data-Centric AI paradigm, the model performance is enhanced without altering the model architecture, as evidenced by real-world and benchmark dataset demonstrations. With the advancements of large language models (LLM), it has become increasingly feasible to generate high-quality synthetic data, while considering the need to construct fully synthetic datasets for real-world data containing numerous personal information. However, in-depth validation of the solely synthetic data setting has yet to be conducted, despite the increased possibility of models trained exclusively on fully synthetic data emerging in the future. Therefore, we examined the question, "Do data quality control techniques (known to positively impact data-centric AI) consistently aid models trained exclusively on synthetic datasets? ". To explore this query, we performed detailed analyses using synthetic datasets generated for speech recognition postprocessing using the BackTranScription (BTS) approach. Our study primarily addressed the potential adverse effects of data quality control measures (e.g., noise injection and balanced data) and training strategies in the context of synthetic-only experiments. As a result of the experiment, we observed the negative effect that the data-centric methodology drops by a maximum of 44.03 points in the fully synthetic data setting.

INDEX TERMS Korean grammatical error correction, synthetic data, noise injection, balanced data.

I. INTRODUCTION

In recent years, the natural language processing (NLP) field has witnessed a surge in practical research on rapidly enhancing model performance without modifying the model in real-world scenarios [1], [2], [3]. This trend has resulted in an expansion of data-centric AI research. Numerous studies are focusing on improving the model performance without model modifications [4]. The research in this domain encompasses a variety of topics such as data management for generating training data [5], data quality control efforts like filtering [6], noise injection [7], [8], as well as data augmentation techniques [4], [9].

Another data management method that is attracting increasing attention is the use of synthetic data [10], particularly as large language models (LLMs) continue to evolve. The emergence of models such as GPT-3 [11], ChatGPT,[1] and LaMDA [12] has enabled the generation of high-quality synthetic data. Consequently, the likelihood of developing models trained solely on synthetic data has increased steadily.

Table 1 presents examples of ChatGPT performing natural language understanding tasks such as morphological analysis

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

[1]https://chat.openai.com/

TABLE 1. Example of generating ChatGPT synthetic data. “User” and “ChatGPT” means user query and ChatGPT system response, respectively.

User:	Analyze ’I know that your company’s system is excellent.’ morphologically
ChatGPT:	Analyzing "I know that your company’s system is excellent." morphologically, we have: - I: Pronoun - know: Verb - that: Subordinating conjunction - your: Possessive adjective - company: Noun - ’s: Possessive ending - system: Noun - is: Verb - excellent: Adjective - .: Punctuation ...
User:	Analyze ’I know that your company’s system is excellent.’ parsing
ChatGPT:	Analyzing "I know that your company’s system is excellent." in terms of parsing, we have: - Sentence: Simple declarative sentence - Subject: "I" - Predicate: "know that your company’s system is excellent." - Verb: "know" - Complement: "that your company’s system is excellent." - Subordinating conjunction: "that" - Noun clause: - Possessive adjective: "your" - Noun: "company" - Possessive ending: "’s" - Noun: "system" - Verb: "is" - Adjective: "excellent" ...
User:	"I know that your company’s system is excellent." translate it into Korean
ChatGPT:	"I know that your company’s system is excellent."는 한국어로 다음과 같이 번역할 수 있습니다: "저는 귀사의 시스템이 훌륭하다는 걸 알고 있습니다."
User:	"I know that your company’s cystem is excellent." correct the grammar
ChatGPT:	"I know that your company’s cystem is excellent." contains a spelling error, as "cystem" should be "system." To correct the sentence, you can write: "I know that your company’s system is excellent." This sentence is grammatically correct and expresses the idea that the speaker is aware that the system belonging to the company in question is of high quality.

and syntactic parsing, as well as machine translation and grammar correction. The sentences generated by ChatGPT can be reused as synthetic data. This demonstrates a direct link between ChatGPT’s generative capabilities and synthetic data generation [13].

Although previous research has indicated a positive impact on model performance, a question remains: have models trained exclusively on synthetic data been validated? Previous studies were predominantly based on human-constructed or publicly available datasets [4], [14]. For Data-Centric AI, the mainstream research has focused on efficiently generating high-quality synthetic data [15] and humanlike data [16]. However, the in-depth validations of models trained solely on synthetic data are relatively few.

In real-world scenarios, obtaining certain data types such as documents containing personal information (e.g., identification cards, medical receipts, and prescription receipts), voice phishing-related audio data for detecting phishing attempts, and documents with substantial defense information can be challenging [17], [18], [19]. Consequently, researchers frequently rely on synthetic data to generate these datasets. Furthermore, with the increase in the number of synthetic data studies based on self-supervised learning [20], [21], [22] and the demonstrated effectiveness of exponentially increasing the amount of data through scaling laws [23], [24], [25], there is a pressing need for in-depth validation of models trained exclusively on synthetic data.

Due to the advancements in LLM, the ability to generate high-quality synthetic data and the growing demand for synthetic data in real-world, it is anticipated that we will rely on models trained on fully synthetic data in the future. However, there is a lack of research on whether the

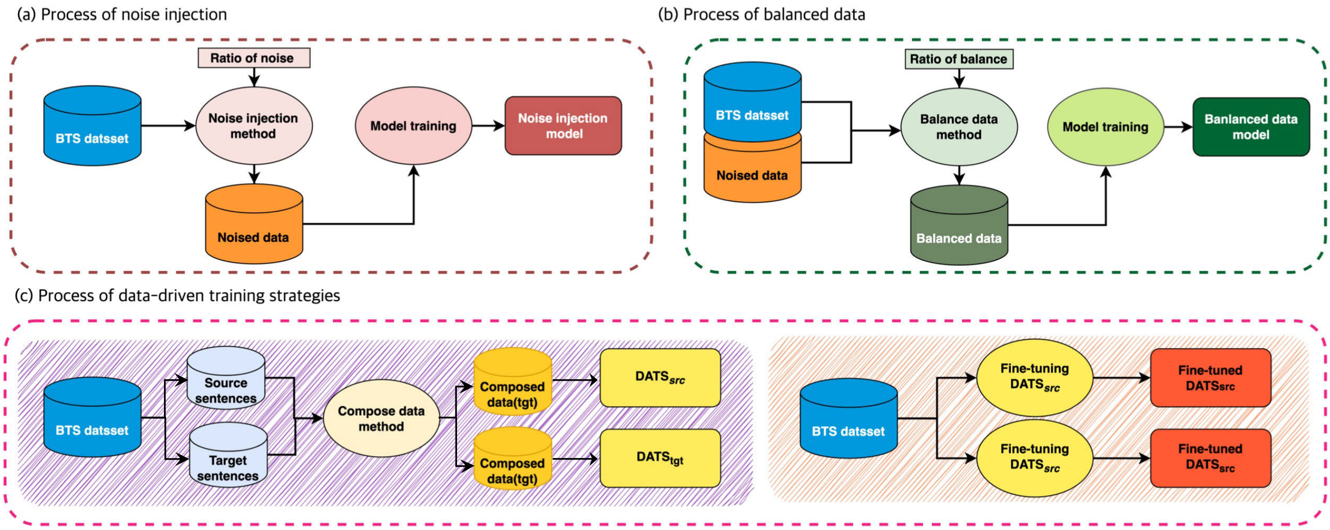

FIGURE 1. Process of research question. (a) indicates the process of noise injection. (b) indicates the process of balanced data. (c) indicates the process of data-driven training strategies. The purple background represents the process of generating the DATS_{src}, DATS_{tgt}. The orange background shows the process of fine-tuning with the each DATS model BTS data.

current methodologies remain equally effective in this exclusive synthetic data setting. In response to this, we conduct pre-validation as a measure to address this trend. Therefore, we investigated whether a data-centric approach can exert a similar positive impact on models trained only on synthetic data. To this end, we revisited and analyzed the research on synthetic data. To the best of our knowledge, our research represents the first comprehensive study that extensively compares various data-centric methodologies known to be effective in real-world datasets within a fully synthetic data setting. We conducted experiments based on a speech recognition post-processing (GEC) task. For performance validation using only synthetic data, we adopted a model based on the recently proposed back transcription (BTS) methodology [26] to generate synthetic data in GEC. Using this model as a foundation, we compared and analyzed the positive and negative impacts on the performance by applying data quality control measures advocated in Data-Centric AI (i.e., noise injection [27] and balanced data [28], [29]), as well as various data-driven training strategies.We highlight the necessity for conducting research on utilizing data quality control in the context of exclusive synthetic data settings through comparative analysis.

Section II provides an explanation for the necessity of research on exclusive synthetic data by showcasing the advancements in LLM and presenting studies in domains that require synthetic data. Section III describes the design of three experiments aimed at validating whether Data-Centric methodologies exhibit similar effects in both the exclusive synthetic data and real-world data settings. Section IV outlines the experimental setup, while Section V-A analyzes the results when applying noise injection (perturbation) methods. Section V-B compares the results when employing balanced data methods, and Section V-C examines the results when implementing data-driven training strategies.

The contributions of this study are as follows.

- We predict the emergence of models trained solely on synthetic data due to the advancements in LLM and the presence of personally identifiable information in real-world datasets, and we conduct pre-experiments in anticipation of this trend.
- We perform comparative research to investigate whether data-centric approaches that have shown positive effects in the real-data environment also exhibit the same effects when applied to models trained solely on synthetic data.
- The experimental results demonstrate that the same methodology has varying impacts on real-world and synthetic data, thereby indicating the need for research on utilizing data quality control in the only synthetic data setting.

II. WHY SYNTHETIC DATA?

The objective of applying generative AI (GAI) techniques (e.g., ChatGPT) in data generation is to make data more efficient and accessible. This enables the generation of high-quality synthetic data at a faster pace [30]. The GAI has been utilized in various fields because of its capability to identify the intent of the instructions provided and generate appropriate content accordingly [31], [32]. With the increase in the data volume and computational capability, several studies have attempted to utilize new technologies by applying GAI algorithms. ChatGPT [1] is a notable example of this.

In November 2022, OpenAI released a novel AI-powered chatbot called ChatGPT. It has attracted considerable attention within the research community. A study determined that participants could differentiate between ChatGPT-generated abstracts of scientific papers and those authored by humans 68% of the time [33]. Moreover, the chatbot attained the

TABLE 2. Example of noise injection according to noise type.

Correct sentence	저는 귀사의 시스템이 훌륭하다는 걸 알고 있습니다. (I know that your company's system is excellent.)
Example of separation error	저는 귀ㅅ ㅏ의 시스템이 훌륭하다는 걸 알고 있습니다. (I know that your com pany's system is excellent.)
Example of vowel transformation error	저는 귀사의 시스템이 훌륭하다는 걸 알교 있습니다. (I kniw that your company's system is excellent.)
Example of pronunciation error	저는 귀사의 시스테미 훌륭하다는 걸 알고 있습니다. (I know that your company's siseutem is excellent.)
Example of punctuation errors	저는 귀사의 시스템.이 훌륭하다는 걸 알고 있습니다. (I know that your company's system. is excellent.)
Example of foreign language transformation error	저는 귀ㅅr의 시스템이 훌륭하다는 걸 알고 있습니다. (I know that you ㅏ company's system is excellent.)
Example of neologism error	저는 귀사의 시스템ㅇ! 훌륭하다는 걸 알고 있습니다. (I know that your company's system !s excellent.)

passing threshold with 60% accuracy on US medical licensing examinations even without specialized human input [34]. These observations indicate that synthetic data produced by generative AI models may progressively gain substantial influence, potentially supplanting real data in specific domains.

Also, synthetic data are effective when real-world data are significantly small or when sensitive data are unavailable. For example, cybersecurity systems require labeled data to identify known malicious activities. However, these are generated manually and seldom released publicly. References [35] and [36] discussed the effectiveness of using synthetic data to evaluate an attacker's capabilities and a system's robustness before system deployment in the simulation for evaluating anomaly detectors.

Detecting small targets in infrared images is important in military systems. However, it is challenging because the non-availability of sufficient structural features may result in false alarms for target discrimination. Reference [37] utilized generative adversarial networks (GAN) [38] to obtain synthetic data without modeling the imaging pipeline or physical world.

Finance is a representative field that is difficult to research owing to the personal information problems. Financial data are confidential and contain personally identifiable customer attributes. The sharing of data outside the business for research purposes is limited stringently. Reference [39] emphasized the importance of generating synthetic financial data with attributes identical to those of real data while protecting personal information. The guidelines for this are presented in the laws of various organizations related to education and medical data privacy protection, e.g., the General Data Protection Regulation (GDPR) laws in the European Union [40], and Family Educational Rights and Privacy Act (FERPA) [41] and Health Insurance Portability and Accountability Act (HIPAA) [42] in the United States.

In addition, synthetic data may be effective in few-shot settings. Typically, table-based question answering (QA) systems need to understand both the natural language (NL) question and the table to produce an appropriate response in few-shot settings. Reference [43] hind the disparity between structured and natural language. Training on synthetic NL questions can close this disparity, particularly when limited annotated NL questions are available.

Given the advancements in LLMs such as ChatGPT, it is imperative to emphasize research solely dedicated to synthetic data generation. This is particularly crucial considering the growing demand for high-quality synthetic data across various domains and applications. The development of LLMs has provided an enabling environment for creating synthetic data of exceptional quality, further underscoring the need for focused research in this area.

III. REVISITING SYNTHETIC DATA

We believe that the advancements in LLM have made it easy to generate synthetic data and given the necessity of synthetic data due to privacy concerns, we anticipate a future where models trained solely on synthetic data will be utilized. In this study, the primary objective was to determine whether *"data quality control (known to be advantageous in conventional cases) exerts a similar positive impact when applied to models trained only on synthetic data."* For a more precise inspection, we designed the following three experimental setups:

A. HOW DOES THE STRENGTH OF NOISE INJECTION IMPACT THE MODEL PERFORMANCE?

First, we aim to examine the effects of applying noise injection methods (*i.e.,* , perturbation), a representative approach for data quality control, in models composed only of synthetic data. Figure 1(a) illustrates the noise injection process. The dataset utilized is constructed with parallel sentences consisting of source and target sentence pairs, similar to those in previous BTS studies.

For perturbations in synthetic data, such as BTS, we apply six types of noise (separation, vowel transformation, pronunciation, punctuation, loanword transformation, and neologism errors) to source sentences. **Separation** refers to a case where the consonants and vowels in a character are separated. **Vowel transformation** replaces a vowel within a character with another vowel. **Pronunciation** is a case where each word is altered according to its pronunciation.

We conduct various case studies based on six noise-injection intensities ranging from 0.1 to 1.0 (0.1, 0.2, 0.4, 0.6, 0.8, and 1.0). Noise is inserted according to the proportion of noise set at the word level within each sentence. For example, if the noise ratio was 1.0, noise occurred in each word within a sentence. Using the generated data, the model was trained to obtain a noisy injection model. We analyze the impact on the performance when applying perturbation to an environment with only synthetic data by comparing this model with a baseline that does not undergo noise injection.

B. HOW DOES THE RATIO OF NOISY AND CLEANED TEXT BATCHES AFFECT THE MODEL PERFORMANCE?

Second, we examine the effects of applying the balanced data method to models using only synthetic data. Balanced data refer to a methodology that constructs datasets by intentionally setting the proportions of noisy and clean data for use in training. Clean data represent unprocessed synthetic data without additional modifications such as noise injection. Meanwhile, noisy data refer to the BTS dataset with applied transformation techniques. For example, if the set ratio is 5:5, 50% of the total data would contain noise. The transformation techniques applied to the noised data are the same perturbations as in Figure 1(a). The goal is to observe the effects of varying the proportion of clean BTS data under identical noise settings.

Figure 1(b) illustrates the process of obtaining balanced data. The clean BTS data and noised data generated in (a) are combined according to the ratio of the balance settings. In order to observe the impact of the ratio of noisy data within balanced data, we constructed balanced data using clean and noisy data in five ratios (5:5, 4:6, 3:7, 2:8, and 1:9). For each setting, we train a model using the generated balanced data. By comparing the performance of the baseline model trained on clean BTS data and the models trained on balanced data using the applied method, we analyzed *the impact of different learning approaches based on noise and clean data ratios on the performance of models trained only on synthetic data.*"

C. DO TRAINING STRATEGIES HAVE A POSITIVE IMPACT ON MODEL PERFORMANCE?

Third, we examine the effects of applying various data-driven training strategies to synthetic data. Figure 1(c) illustrates the data-driven training strategy. The synthetic parallel data used here consisted of source sentences containing errors and target sentences without noise. This process demonstrates the generation of data-driven training strategy (DATS) models by constructing various combinations of training strategies.

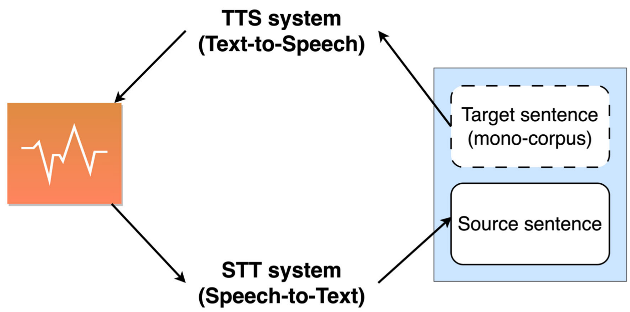

FIGURE 2. Architecture of BackTranScription (BTS).

The purple section in (c) shows the process of constructing parallel corpora using only data with identical characteristics. To achieve this, we separate the parallel corpora with different characteristics and reassembled pairs using sentences with identical training characteristics. That is, we conduct training using source-source sentence pairs from parallel sentences. Conversely, we separate target sentences from parallel sentences and conducted training using target–target sentence pairs. This implies that we train the DATS models using sentence pairs with identical characteristics, both containing errors and those without, resulting in DATS*src* and DATS*tgt* models, respectively. This functions as a type of ablation for the next step, i.e., fine-tuning on synthetic data.

The orange section in (c) illustrates the process of applying additional fine-tuning to each DATS model using synthetic data. This involves post-training, where the trained DATS models are fine-tuned further using synthetic data composed of source–target sentence pairs. That is, this strategy sequentially trains data with both identical and different characteristics. Through this process, we obtain fine-tuned DATS*src* and fine-tuned DATS*tgt* models.

We analyze the impact of different Data-Centric AI learning strategies on the model performance by comparing a baseline model with a model trained using two data-driven learning strategies.

IV. EXPERIMENTAL SETTINGS

A. SYNTHETIC DATA

We generate synthetic datasets for training data using BTS [26]. Figure 2 shows the architecture for synthesizing monolingual corpus into a parallel corpus using BTS. Training a post-processor (GEC) requires parallel corpora created by human annotators, which are expensive and not scalable. To alleviate this problem, BTS combines text-to-speech (TTS) technology and speech-to-text (STT) technology to generate GEC task synthesized data for speech recognition post-processor. The synthesized voice data generated by TTS undergoes text-to-speech conversion through Navers CLOVA Speech Recognition (CSR) API. This API employs the same model utilized for Navers Voice Recognition Notes and Searches, requiring less than 120 hours and 72 hours for conversion, respectively. The voice data is synthesized using the

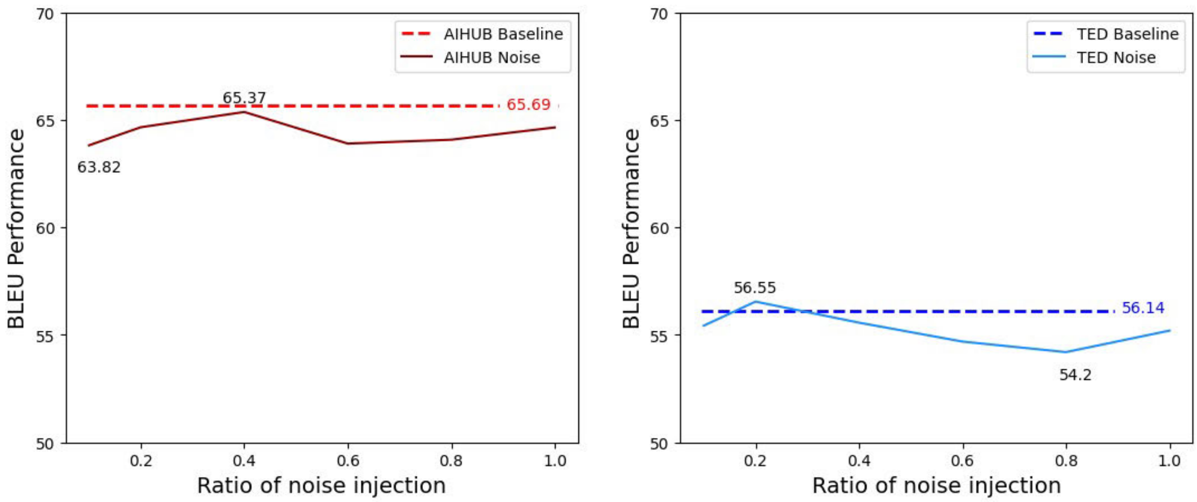

FIGURE 3. Experimental results of noise injection. (a) is the result of inserting noise into synthetic data of AIHUB. (b) is the result of inserting noise into synthetic data of TED. Note that the x-axis indicates the strength of noise injection.

TABLE 3. Statistics of the raw dataset in synthetic. Each value represents the number of sentences included in the data. Dev. indicates the development set.

Dataset	Train	Dev.	Test	Total
AI-HUB	92,000	5,000	3,000	100,000
TED	119,883	5,000	3,000	127,883

WaveNet model, the same model employed by Google Assistant, Google Search, and Google Translation, with conversion times of less than 36 hours and 24 hours, respectively. The process of building a parallel corpus involves the following steps: 1) crawling the pre-built mono corpus in a convenient manner; 2) transformation into speech using TTS; and 3) outputting the converted result as text using STT. Although BTS may not encompass the field of synthetic data, BTS is a simple and efficient methodology for generating synthetic data. Thus, we use it for experiments. We employ the AI-HUB [44], which are representative Korean data platform, and the TED Korean dataset,[2] the same datasets used in prior BTS studies, as raw data for generating BTS-based synthetic data. AI-HUB and TED datasets, being well-curated and openly available, are highly suitable for utilization as raw monolingual corpora due to their ease of access. We use the BTS methodology to create synthetic parallel data from raw monolingual corpus. In addition, since the existing BTS research was also conducted in Korean, this experiment also performs based on Korean for a fair evaluation.

Table 3 shows the data statistics used in the experiment. We utilize 92,000 sentences from AI-HUB data and 119,883 sentences from TED's Korean Transcript data to generate BTS-based synthetic data. In order to analyze the impact of synthetic data on the model more accurately, we maximize the quantity of training data within the available resources. These data are used as raw data for BTS, transformation into speech using TTS and outputting the converted result as text using STT. In other words, TST (Text-to-Speech-to-Text) technology is utilized to generate a pseudo-parallel corpus, which comprises synthetic data of (TTS result, original monolingual corpus text) including generated noise during each phase of the process.

[2] https://www.ted.com/talks?language=ko

B. IMPLEMENTATION DETAILS

In order to focus on the performance of the methodology rather than the model itself, we employed the vanilla transformer [45] for model training. The hyperparameters set to the values in [45]. Fairseq [46] is used for the implementation. For subword tokenization, we utilize SentencePiece [47] and set the vocabulary size to 50,000 words. We evaluate the BTS-based synthetic data models BLEU [48] and GLEU [49]. A higher value for BLEU and GLEU indicates better performance. These metrics are identical to those used in previous BTS studies.

V. EXPERIMENTAL RESULTS

A. RESULTS FOR QUESTION 1: NOISE INJECTION

The experimental results applying noise injection (perturbation) methods are shown in Figure 3. (a) Synthetic data generated based on AIHUB and (b) TED. The noise injection ratio refers to the probability that tokens in sentences would be inserted with noise. The baseline performances are 65.69 and 56.14 when trained with synthetic data from AIHUB and TED, respectively. In most cases, the performance deteriorated, reporting the lowest performance of 63.82, 54.2 when the noise ratio is 0.1, 0.8 and the highest performance of 65.37, 56.55 for the noise ratio 0.4, 0.2. In particular, the performance of TED outperforms the baseline by a small margin when the noise ratio is set to 0.2.

Compared with models trained on data using perturbation methods, the performance tend to improve when the noise

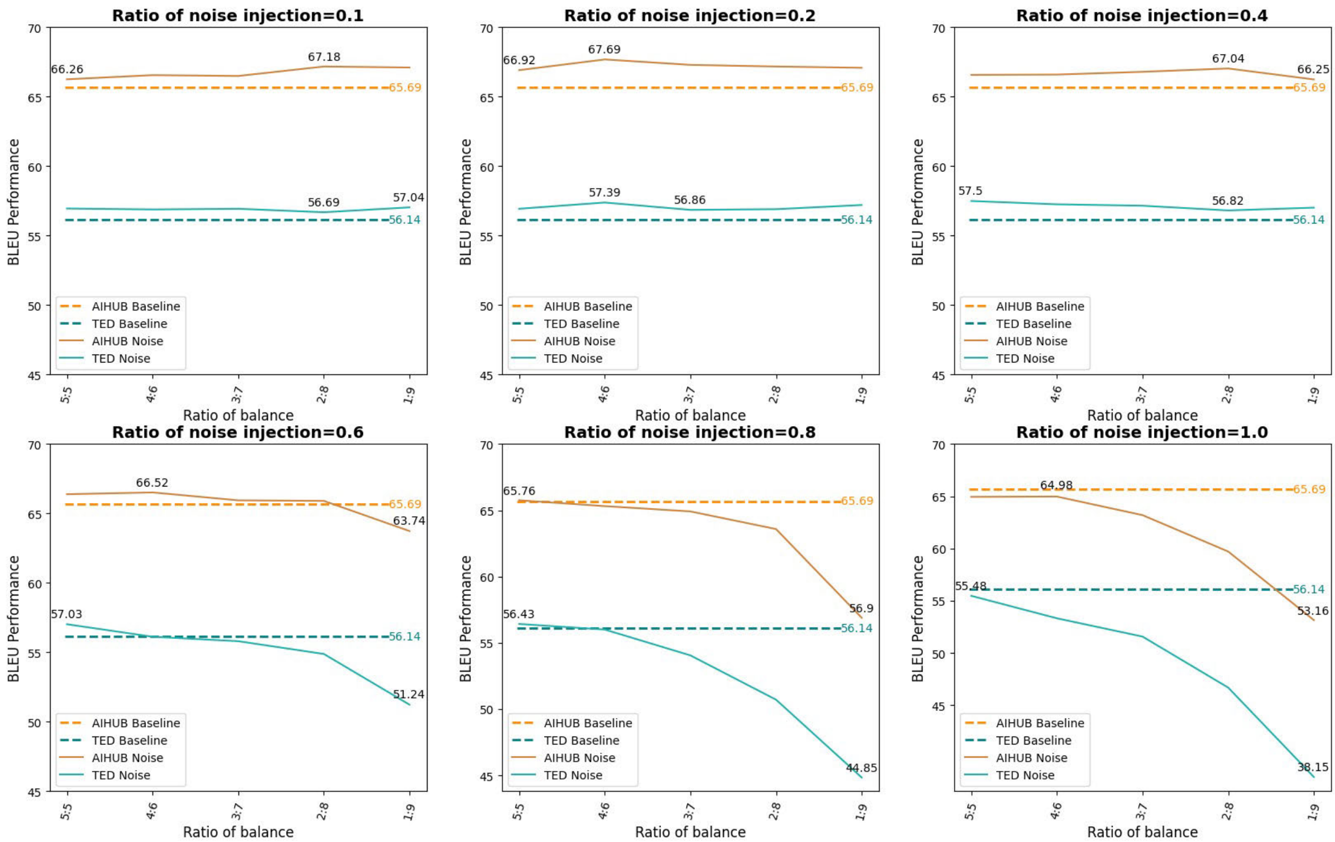

FIGURE 4. **Experimental results of balanced data. Note that the x-axis indicates the ratio as (clean:noise).**

ratio was low. However, the strong negative impact ratios are 0.1, 0.8, for AIHUB and TED, respectively, which cannot specify the tendency. We infer that the ratio causing negative impact is inconsistent. Perturbation methods in synthetic-data-only settings have a negative impact on performance and are out of control. *We conclude that the positive impact of the perturbation methods demonstrated in numerous previous studies show a negative impact on the synthetic-data-only setting.* This indicates the need for research in environments that consider only synthetic data, rather than relying on traditional methodologies based on real-world data settings.

B. RESULTS FOR QUESTION 2: BALANCED DATA

The second set of experimental results, including the application of the balanced data methods, is shown in Figure 4. We provide a ratio according to the balance setting ratio to clean synthetic data that do not require further processing and noisy data to which perturbation methods have been applied. We then train the model with the data. The clean-to-noise data ratio is set to five as mentioned earlier. The noise injection ratio is applied identically to Figure 3. The result reveals performance degradation when the noise injection ratio is over 0.6. Especially, AIHUB results deteriorate by 19.07% (53.16) compared to the baseline when the ratio of clean and noise data is 1:9 and the noise ratio is 1.0. The TED results consistently showed no progress at 32.04% (38.15) below the baseline. We find that the performance gap occurs when the noise injection ratio increases and the ratio of clean data and noise data is skewed towards the noisy data.

Meanwhile, progress is achieved in the noise injection ratio 0.1–0.4 regardless of the ratio of balance setting. The performance of AIHUB tended to improve when the clean data ratio is smaller in the data with the ratios 0.1, 0.4. We conjecture that when the noise ratio is low, the model is robust regardless of the ratio of the balance setting and an appropriate amount of noise is required. However, the synthetic-data-only environment shows also a positive impact as well as a negative, unlike the real-world environment. This implies that it is necessary to determine an appropriate balance ratio setting at a low injection ratio to effectively use data quality control.

The experimental results show ***negative effect*** as the ratio of balance increases in a situation where the ratio of noise injection is high. Namely, we confirm that *in an environment composed of only synthetic data, data quality control may have a less positive effect on the model.* Through this, data quality control, which is known to have a positive effect on real-world data-based models, *we conclude that a positive effect is not always guaranteed in an environment made up of only synthetic data.* These results imply that real-world data and synthetic data have distinctly different characteristics and must be dealt with separately. Therefore, we suggest

TABLE 4. Experiment results (%) of different models on AIHUB and TED datasets.

Model	AIHUB		TED	
	BLEU	GLEU	BLEU	GLEU
Base	39.68	-	46.09	-
BTS baseline	65.69	55.70	56.14	49.48
DATS$_{src}$	45.82 (-19.87)	17.64 (-38.06)	39.27 (-16.87)	11.17 (-38.31)
DATS$_{tgt}$	30.98 (-34.71)	27.88 (-27.82)	39.27 (-16.87)	34.32 (-15.16)
Fine-tuned DATS$_{src}$	44.90 (-20.79)	37.94 (-17.76)	39.23 (-16.91)	34.20 (-15.28)
Fine-tuned DATS$_{tgt}$	21.66 (-44.03)	19.89 (-35.81)	38.83 (-17.31)	33.86 (-15.62)

that beyond effective synthetic data generation, which is the main focus of the existing Data-Centric AI, research must be conducted to ensure that models using synthetic data can produce sufficient performance.

C. RESULTS FOR QUESTION 3: DATA-DRIVEN TRAINING STRATEGIES

The experimental results of applying the training strategy are presented in Table 4. The DATS$_{src}$ and DATS$_{tgt}$ models are the results of learning only with data exhibiting identical characteristics. In the case of DATS$_{src}$ (which are trained by constructing data only with error sentences), AIHUB and TED records 45.82 and 39.27 BLEU, respectively. These represent reductions by -19.87 and -16.87 points, respectively. In the case of the DATS$_{tgt}$ model, AIHUB reduces by -34.71 points to 30.98, and TED degraded by -16.87 points to 39.27. Although the performance of DATS$_{src}$ shows a smaller decrease than that of DATS$_{tgt}$, both models exhibit a reduction from the baseline.

The fine-tuned DATS$_{src}$ and fine-tuned DATS$_{tgt}$ models indicate the results of applying a method of sequentially training the data with identical and different characteristics. We fine-tune the BTS data in addition to the previous model and present the results. In the case of fine-tuned DATS$_{src}$, the BLEU score of AIHUB is 44.9 (-20.79 points lower than the baseline) and that of TED is 39.23 (-16.91 BLEU lower). The fine-tuned DATS$_{tgt}$ model scores -44.03 points lower than the baseline at 21.66 for AIHUB and -17.31 points lower at 38.83 for TED.

This indicates that even when additional fine-tuning is conducted with the synthesized data, it has an adverse effect on performance improvement. In addition, similar to the scenario before the additional fine-tuning, the performance of the TED-based model is better than that of AIHUB. We infer that the initial model training is important.

We experiment with learning strategies by constructing various cases. However, the pure model without any applied strategy exhibits the best performance. Thus, we verify the occurrence of an adverse effect of applying a data-based learning strategy in the context of synthetic data. This implies that need to go beyond effective synthetic data generation in the traditional Data-Centric AI field and conduct research to ensure that models trained on synthetic data can achieve sufficient performance.

VI. CONCLUSION AND FUTURE WORK

Recently, research has focused on improving the model performance without modifying the model or using synthetic data as an approach. This study investigated the research question of whether data quality control (a Data-Centric AI methodology known to have a positive impact) also yields a positive impact when applied only to models trained on synthetic data. To answer this question, we conducted experiments by applying data quality control techniques such as noise injection (perturbation), balanced data, and various training strategies. The results revealed an adverse impact in models trained only on synthetic data. Notably, when employing the balanced data method, the AIHUB dataset experienced a decline of 19.07% in performance compared to the baseline when the ratio of clean and noise data is 1:9, while the TED dataset consistently exhibited no improvement, performing 32.04% below the baseline. This demonstrated that data-centric methodologies do not necessarily ensure a positive effect depending on the characteristics of the data. This highlights the need for sufficient verification of data-centric methodologies in synthetic data environments and for research to utilize data quality control. We plan to conduct additional analyses of the characteristics of synthetic data environments by applying various data-centric approaches.

ACKNOWLEDGMENT

(Seonmin Koo and Chanjun Park contributed equally to this work.)

REFERENCES

[1] A. E. Goosen, "A system to quantify industrial data quality," Ph.D. dissertation, North-West Univ., Potchefstroom, South Africa, 2019.

[2] M. Tarafdar, C. M. Beath, and J. W. Ross, "Using AI to enhance business operations," *MIT Sloan Manage. Rev.*, vol. 60, no. 4, pp. 1–9, 2019.

[3] S. Laato, M. Tiainen, A. K. M. N. Islam, and M. Mäntymäki, "How to explain AI systems to end users: A systematic literature review and research agenda," *Internet Res.*, vol. 32, no. 7, pp. 1–31, Dec. 2022.

[4] M. Mazumder et al., "DataPerf: Benchmarks for data-centric AI development," 2022, *arXiv:2207.10062*.

[5] E. Choi and C. Park, "DMOps: Data management operation and recipes," 2023, *arXiv:2301.01228*.

[6] P. Koehn, V. Chaudhary, A. El-Kishky, N. Goyal, P.-J. Chen, and F. Guzmán, "Findings of the WMT 2020 shared task on parallel corpus filtering and alignment," in *Proc. 5th Conf. Mach. Transl.*, 2020, pp. 726–742.

[7] S. Sarp, M. Kuzlu, U. Cali, O. Elma, and O. Guler, "Analysis of false data injection impact on AI based solar photovoltaic power generation forecasting," 2021, *arXiv:2110.09948*.

[8] A. Partovyan, V. Nourani, and M. T. Alami, "Noise injection–denoising techniques to improve artificial intelligence-based rainfall–runoff modeling," *Water Resour. Eng.*, vol. 11, no. 36, pp. 81–94, 2018.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[10] S. I. Nikolenko, "Synthetic data for deep learning," 2019, *arXiv:1909.11512*.

[11] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[12] R. Thoppilan et al., "LaMDA: Language models for dialog applications," 2022, *arXiv:2201.08239*.

[13] S. Ubani, S. O. Polat, and R. Nielsen, "ZeroShotDataAug: Generating and augmenting training data with ChatGPT," 2023, *arXiv:2304.14334*.

[14] N. Polyzotis and M. Zaharia, "What can data-centric AI learn from data and ML engineering?" 2021, *arXiv:2112.06439*.

[15] C. Park, S. Lee, H. Moon, S. Eo, J. Seo, and H. Lim, "How should human translation coexist with NMT? Efficient tool for building high quality parallel corpus," 2021, *arXiv:2111.00191*.

[16] H. Moon, C. Park, J. Seo, S. Eo, and H. Lim, "An automatic post editing with efficient and simple data generation method," *IEEE Access*, vol. 10, pp. 21032–21040, 2022.

[17] M. Meingast, T. Roosta, and S. Sastry, "Security and privacy issues with health care information technology," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2006, pp. 5453–5458.

[18] D. S. Terzi, R. Terzi, and S. Sagiroglu, "A survey on security and privacy issues in big data," in *Proc. 10th Int. Conf. Internet Technol. Secured Trans. (ICITST)*, Dec. 2015, pp. 202–207.

[19] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet, "The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization," *Comput. Linguistics*, vol. 48, no. 4, pp. 1053–1101, Dec. 2022.

[20] N. Ng, K. Cho, and M. Ghassemi, "SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 1268–1283. [Online]. Available: https://aclanthology.org/2020.emnlp-main.97

[21] D. Ruiter, D. Klakow, J. van Genabith, and C. España-Bonet, "Integrating unsupervised data generation into self-supervised neural machine translation for low-resource languages," in *Proc. Mach. Transl. Summit XVIII, Res. Track*. Washington, DC, USA: Association for Machine Translation in the Americas, Aug. 2021, pp. 76–91. [Online]. Available: https://aclanthology.org/2021.mtsummit-research.7

[22] Z. Gan, H. Xu, and H. Zan, "Self-supervised curriculum learning for spelling error correction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3487–3494.

[23] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[24] T. E. Raghunathan, "Synthetic data," *Annu. Rev. Statist. Appl.*, vol. 8, no. 6, pp. 129–140, 2021.

[25] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020, *arXiv:2001.08361*.

[26] C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo, and H. Lim, "BTS: Back transcription for speech-to-text post-processor using text-to-speech-to-text," in *Proc. 8th Workshop Asian Transl. (WAT)*, 2021, pp. 106–116.

[27] M. Ivanovs, R. Kadikis, and K. Ozols, "Perturbation-based methods for explaining deep neural networks: A survey," *Pattern Recognit. Lett.*, vol. 150, pp. 228–234, Oct. 2021.

[28] C. Park, W.-Y. Go, S. Eo, H. Moon, S. Lee, and H. Lim, "Mimicking infants' bilingual language acquisition for domain specialized neural machine translation," *IEEE Access*, vol. 10, pp. 38684–38693, 2022.

[29] L. Chen, S. Wan, and L. Dou, "Improving diagnostic performance of high-voltage circuit breakers on imbalanced data using an oversampling method," *IEEE Trans. Power Del.*, vol. 37, no. 4, pp. 2704–2716, Aug. 2022.

[30] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," 2023, *arXiv:2303.04226*.

[31] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539–559, Jan. 2023.

[32] P. Pu Liang, A. Zadeh, and L.-P. Morency, "Foundations and trends in multimodal machine learning: Principles, challenges, and open questions," 2022, *arXiv:2209.03430*.

[33] H. Else, "Abstracts written by ChatGPT fool scientists," *Nature*, vol. 613, no. 7944, p. 423, Jan. 2023.

[34] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digit. Health*, vol. 2, no. 2, Feb. 2023, Art. no. e0000198.

[35] H. Ringberg, M. Roughan, and J. Rexford, "The need for simulation in evaluating anomaly detectors," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 1, pp. 55–59, Jan. 2008.

[36] S. Abt and H. Baier, "A plea for utilising synthetic data when performing machine learning based cyber-security experiments," in *Proc. Workshop Artif. Intell. Secur. Workshop*, Nov. 2014, pp. 37–45.

[37] J.-H. Kim and Y. Hwang, "GAN-based synthetic data augmentation for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002512.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[39] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, "Generating synthetic data in finance: Opportunities, challenges and pitfalls," in *Proc. 1st ACM Int. Conf. AI Finance*, 2020, pp. 1–8.

[40] C. J. Hoofnagle, B. van der Sloot, and F. Z. Borgesius, "The European Union general data protection regulation: What it is and what it means," *Inf. Commun. Technol. Law*, vol. 28, no. 1, pp. 65–98, Jan. 2019.

[41] T. Gilbert, "Family educational rights and privacy act (FERPA)," *J. Empirical Res. Hum. Res. Ethics*, vol. 2, p. 101, Nov. 2007.

[42] I. G. Cohen and M. M. Mello, "HIPAA and protecting health information in the 21st century," *J. Amer. Med. Assoc.*, vol. 320, no. 3, pp. 231–232, 2018.

[43] Z. Jiang, Y. Mao, P. He, G. Neubig, and W. Chen, "OmniTab: Pretraining with natural and synthetic data for few-shot table-based question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Seattle, WA, USA: Association for Computational Linguistics, Jul. 2022, pp. 932–942. [Online]. Available: https://aclanthology.org/2022.naacl-main.68

[44] C. Park, M. Shim, S. Eo, S. Lee, J. Seo, H. Moon, and H. Lim, "Empirical analysis of parallel corpora and in-depth analysis using LIWC," *Appl. Sci.*, vol. 12, no. 11, p. 5545, May 2022.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[46] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "Fairseq: A fast, extensible toolkit for sequence modeling," 2019, *arXiv:1904.01038*.

[47] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," 2018, *arXiv:1808.06226*.

[48] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.

[49] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "Ground truth for grammatical error correction metrics," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 588–593.

SEONMIN KOO received the B.S. degree from the Department of Computer Science and Engineering, Konkuk University, Seoul, South Korea, in 2022, where she is currently pursuing the Ph.D. degree in computer science and engineering from Korea University, Seoul. She is currently under an integrated master's and Ph.D. courses. She is a part of the Natural Language Processing and Artificial Intelligence Laboratory Team. Her research interests include machine translation and knowledge-based populations.

CHANJUN PARK received the Ph.D. degree in computer science and engineering from Korea University, Seoul, South Korea, in August 2023. He is currently working as an Technical Leader (TL) at Upstage LLM Team. From 2018 to 2019, he worked at SYSTRAN as a Research Engineer. He is the Founder and Leader of the KU-NMT Group, and has received the Naver Ph.D. Fellowship in 2021. He served as the Virtual Social Chair at COLING 2022, and is currently serving as the Program Chair for the WiNLP Workshop. He has published more than 150 papers in the field of NLP. His research interests include the Natural Language Processing (NLP), with a focus on Data-Centric AI, Machine Translation and Large Language Model (LLM).

SEOLHWA LEE received the Ph.D. degree in computer science and engineering from Korea University, Seoul, South Korea, in August 2021. She is currently a Postdoctoral Researcher with the Technical University of Darmstadt, Germany. Her research interests include the intersection of human cognition and natural language processing, privacy-aware NLP, dialogue summarization, dialogue agents, text summarization, and deep learning.

JAEHYUNG SEO received the B.S. degree from the Department of English Language and Literature, Korea University, Seoul, South Korea, in 2020, where he is currently pursuing the Ph.D. degree in computer science and engineering. He is currently under an integrated master's and Ph.D. courses. He is a part of the Natural Language Processing and Artificial Intelligence Laboratory Team. His research interests include language generation and decoding strategies, where he attempts to find inspiration from how humans do so and build a generative model based on common-sense reasoning.

SUGYEONG EO received the B.A. degree in linguistics and cognitive science, language, and technology from the Hankuk University of Foreign Studies, Yongin, South Korea, in 2020. She is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul, South Korea. She is a member of the Natural Language Processing and Artificial Intelligence Laboratory. Her research interests include neural machine translation, quality estimation, and question generation.

HYEONSEOK MOON received the B.S. degree from the Department of English Language and Literature, Korea University, Seoul, South Korea, in 2021, where he is currently pursuing the Ph.D. degree in computer science and engineering. He is currently under an integrated master's and Ph.D. courses. He is a part of the Natural Language Processing and Artificial Intelligence Laboratory Team. His research interests include natural language understanding and neural machine translation.

HEUISEOK LIM received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1992, 1994, and 1997, respectively. He is currently a Professor with the Department of Computer Science and Engineering, Korea University. His research interests include natural language processing, machine learning, and artificial intelligence.