ICGHIT

International Conference on
Green and Human Information Technology 2025

ICGHIT 2025

Jan.15 – 17, 2025
Nha Trang, Vietnam

Proceedings

5, 2024
am

GHIT IEIE SsangYong Information & Communications Corp.

http://icghit.org/

ISSN : 2466-121X

Topic Classification Training Model with Automatic Textual Data Transformation

Jinmo Yang*, Janghwan Kim†, Chaeyun Seo‡, Kidu Kim§, and R. Young Chul Kim¶
*†‡SE Lab, Hongik University, Sejong Korea
§Telecommunications Technology Association
¶SE Lab, Hongik University, Sejong Korea,
Emails: *yjmd2222@g.hongik.ac.kr, {†lentoconstante, ‡chaeyun}@hongik.ac.kr, §kdkim@tta.or.kr, ¶bob@hongik.ac.kr

***Abstract*—Large language models (LLMs) have gained significant popularity due to their competency across various domains and tasks. While fine-tuning LLMs enhances performance for any domains and tasks, the potential in the field of linguistics—where identical contexts are expressed in diverse syntax forms—is rarely considered. In this study, we propose the mechanism for fine-tuning LLMs with datasets containing the same context but different syntax. Continuing from our previous work—where we have augmented the Korean Language Understanding Evaluation (KLUE) topic classification (TC) dataset of 45,678 sentences into four syntax forms—we manually created 558 sentence sets in addition. Using the latter dataset, we trained a random forest model, achieving an f1-score of 0.984, significantly outperforming the XLM-R-large model's 0.861 on the original TC dataset. Challenges remain, including combining manually created datasets with the augmented data, conducting an ablation study to assess syntax combinations, and addressing inconsistency in context across syntax types. Nevertheless, this work lays the foundation for further exploration into syntax-aware fine-tuning of LLMs and their applications in any specialized domains and tasks.**

Keywords—data augmentation, f1-score, natural language processing, prompt engineering

I. Introduction

The popularity of large language models (LLMs) is increasing ever since OpenAI's ChatGPT3 was released in 2022 [1]. With the development of the AI and the training techniques, LLMs are becoming more capable of simple questioning and answering, doing minute tasks for the users' convenience, and creating business ideas for profit [2, 3]. This is possible as the models have been trained on huge data of different variety during various learning phases [4]. And to be more effective and competent in specific domains or tasks, the LLMs are fine-tuned with high-quality datasets of the specific domains or task instructions [5]. Now, when the focus is moved to training a model in the domain of specific language, the general acceptance is that simple sentences are understood better than compound, complex, or colloquial sentences. However, constructing datasets of sentences of the exactly the same context but different syntax is mostly ignored. LLMs perform better in a specific domain when they are fine-tuned with the domain-specific dataset; therefore, it is wise that the domain-specific dataset is further preprocessed and multiplied to different syntax to also overlap with the linguistic domain. Therefore, we present a topic classification model training mechanism with automatic textual data transformation. The rest of the paper is organized as follows. Section 2 mentions background research and related works. Section 3 presents our current progress. Finally, Section 4 mentions our conclusion.

II. Background Research and Related Works

There are efforts to evaluate the performance of LLMs with syntactical influence. Kim *et al.* [6] created their own Syntactically Incomplete Korean (SIKO) dataset to evaluate the performance of an LLM in Korean language proficiency if syntax deviates from the formal form. The results showed some improvements from augmentation with syntax deviation. However, their experiment did not include the case with augmented data and original data combined to make a dataset size of original multiplied by two, thereby not accounting the case of the same context and different syntax.

Our previous work augmented original Korean Language Understanding Evaluation (KLUE) topic classification (KLUE) [8] dataset into four datasets of different sentence types: simple, compound, complex, and colloquial sentences [7]. The method for augmentation was prompt-engineering an LLM with the domain knowledge of Korean linguistics, and then giving it original TC data for augmentation. We have used few engineering techniques for high quality datasets and obtained four sets of 45,678 sentence sets in simple, compound, complex, and colloquial forms. Fig. 1 shows the prompt-engineering and data augmentation (transformation) into data collection.

III. Topic Classification Training Model

This paper mostly focuses on the training. The training dataset explanation and the training process are as follows.

A. Training Dataset Explanation

The previously proposed training dataset was automatically transformed from the original data of KLUE TC benchmark dataset [7]. The original text datapoints are transformed into the sentences of simple, compound, complex, and colloquial types. The resulting dataset consists of four transformed sub-datasets of 45,678 sentences each, totaling 182,712 sentences. At the same time, manual sentence creation is being conducted. Currently, there are 558 manually created sentence sets in the four types total.

Presently, the memory utilization on our laboratory's workstation is yet to be optimized; therefore, only the smaller, manually created sentences are used for training. Table 1 shows a subset of the manually created sentence sets.

B. Model Training

The model for training with the sentences of diverse syntax is preferably one of the current state-of-the-art LLMs. The leaderboard of KLUE shows that the XLM-R-large model placed the first in TC with the score of 86.06 [9]. Therefore, this model is suitable for training for the highest score.

As mentioned before, the memory utilization for inferencing is not yet optimized. Therefore, a model of a

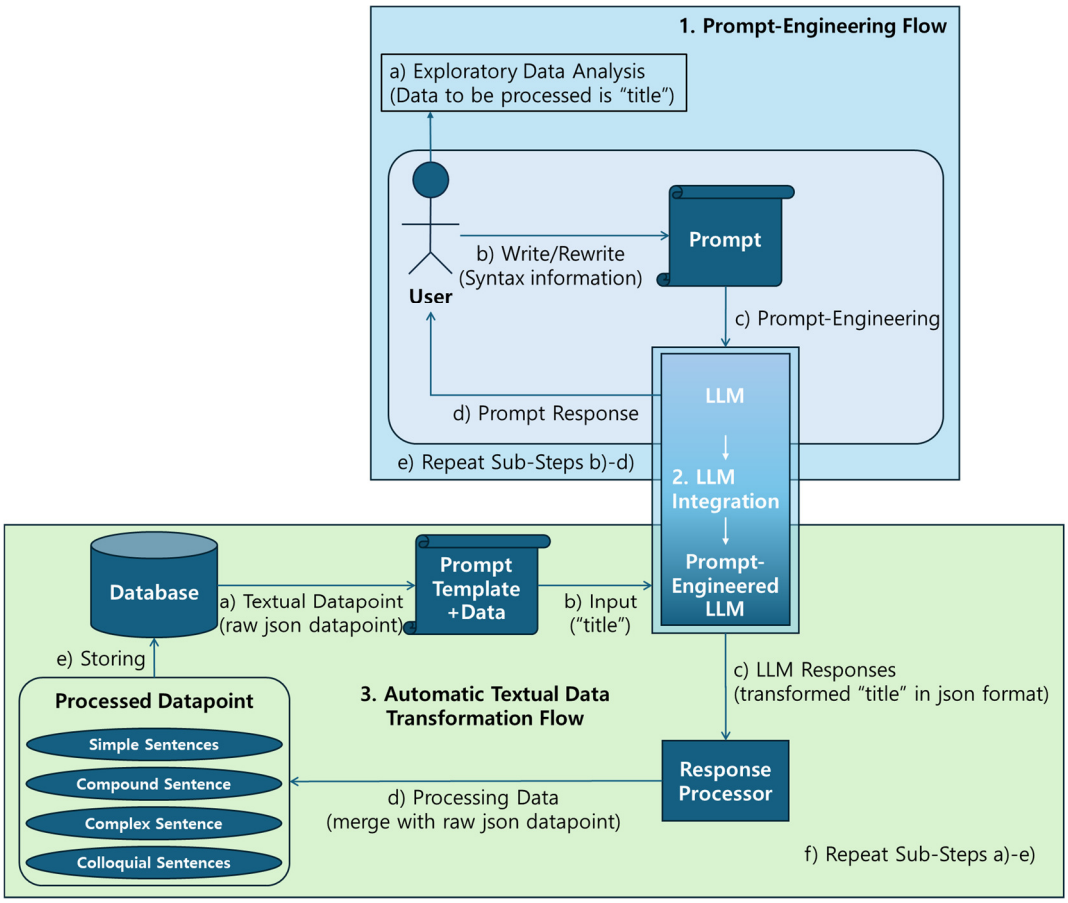

Fig. 1. Automatic Textual Data Transformation

TABLE I. MANUALLY CREATED SENTENCES EXAMPLE

Row	Formal			Informal
	Simple	*Compound*	*Complex*	*Colloquial*
0	사무실 중앙에 테이블이 있다	사무실 한가운데 테이블이 놓여있고, 테이블 중앙에 꽃병을 놓았다.	테이블 한가운데가 비어 보여서, 나는 그곳에 테이블을 놓았다.	테이블 한가운데에 테이블이 놓여있다.
1	의자의 등받이는 검은색 가죽이다.	의자의 등받이가 검은색 가죽으로 되어있었지만, 의자 등받이가 찢어졌다.	내가 사무실에 도착했을 때, 의자의 등받이는 검은색 가죽으로 되어있었다.	의자의 등받이가 검은색 가죽으로 되어있다.

smaller size is needed. Presently, the random forest classifier model is chosen as the training model [10].

Fig. 2 shows the process of the training. The steps are explained as follows:

1. Load data: Manually created data is loaded into a DataFrame (table).
2. Combine sentences: Sentences of each type are combined into a single column.
3. Split rows: Data is split into train and test data.
4. Create embeddings: Vector embeddings are created with term frequency-inverse document frequency vectorizing method.
5. Train: The random forest classifier model is trained from the training data.
6. Evaluate: The performance with the test data is evaluated in the form of f1-score.

From the evaluation, an f1-score of 0.984 is obtained. This is a very high score, considering the score of 0.861 with the XLM-R-large model. However, many factors need to be accounted for. 1) The manually created dataset is incompatible with the original or the transformed KLUE TC datasets as the labels do not match. Zero-shot classification need to be considered. 2) Ablation study is likely needed to determine the effects of different combinations of sentence types on the model performance. 3) Some rows of the manually created dataset have different contexts across the sentence types. The sentences need to be assessed and modified.

IV. CONCLUSION

We mention the topic classification training model mechanism with automatic textual data transformation, using manually created data and the random forest classifier model for testing the mechanism. The model achieved a high f1-score of 0.984, compared to 0.861 from XLM-R-large, the first rank model on KLUE TC leaderboard. For the future works, we will combine the manually created dataset with KLUE TC datasets, optimize the workstations for LLM training, and conduct a full ablation study on the different types of sentences.

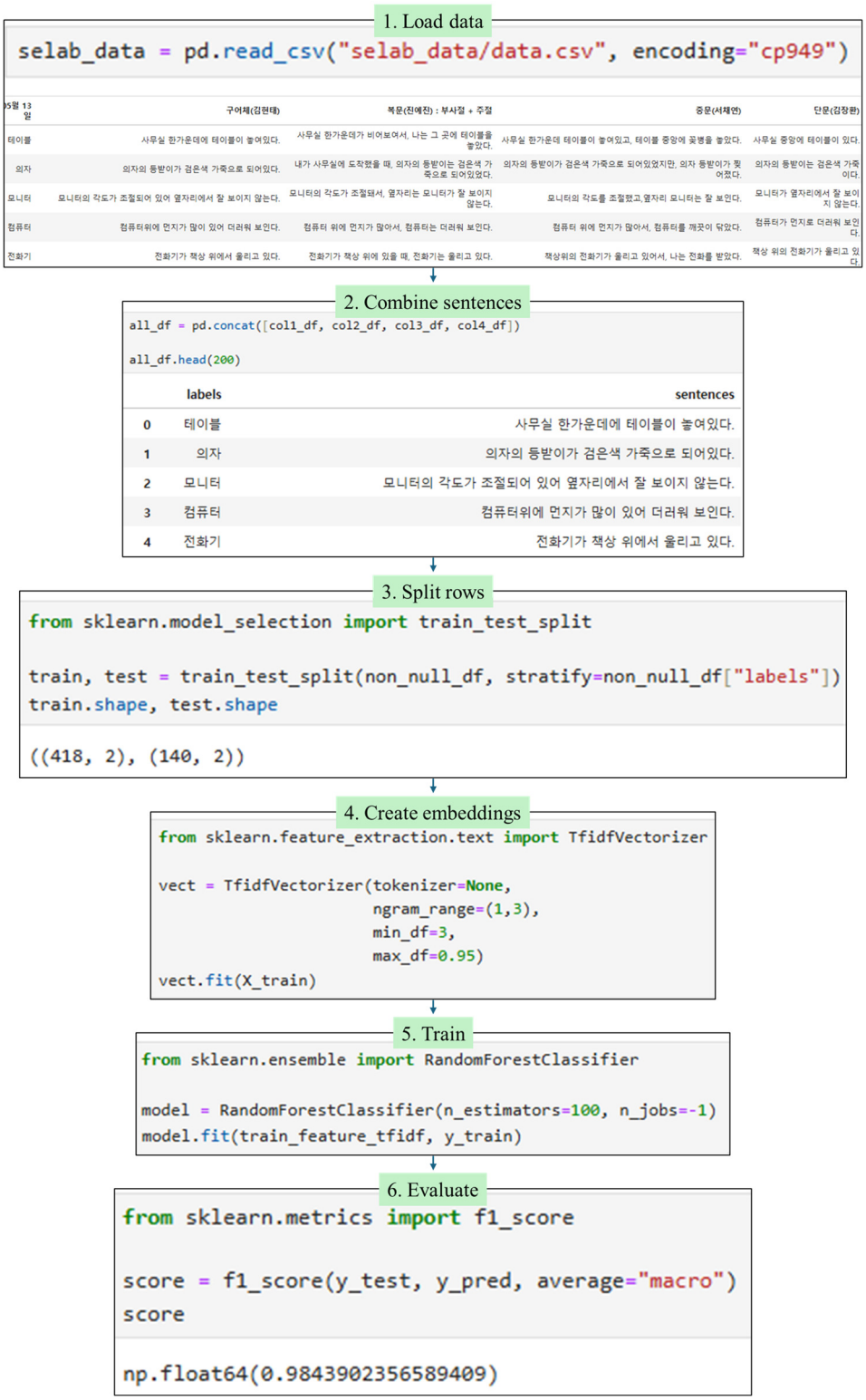

Fig. 2. The training process

ACKNOWLEDGMENT

This research was supported by Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2024 (Project Name: Artificial Intelligence-based User Interactive Storytelling 3D Scene Authoring Technology Development, Project Number: RS-2023-00227917307820870502O1) and National Research Foundation (NRF), Korea, under project BK21 Four.

REFERENCES

[1] OpenAI, "Introducing ChatGPT," OpenAI, https://openai.com/index/chatgpt/, Accessed October 31, 2024.

[2] M. Abdullah, A. Madain, and Y. Jararweh, "ChatGPT: Fundamentals, Applications and Social Impacts," In Proc. of 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2022.

[3] C. Lee, "Design to Improve Educational Competency Using ChatGPT," IJIBC, vol. 16, no. 1, pp. 182-190, 2024.

[4] S. Balasubramaniam, S. Kadry, A. Prasanth, and R. Dhanaraj, Generative AI and LLMs: Natural Language Processing and Generative Adversarial Networks, Berlin, Boston: De Gruyter, 2024.

[5] D. Jang, S. Byun, H. Jo, and H. Shin, "A Comprehensive Korean Instruction Toolkit on 19 Tasks for Fine-Tuning Korean Large Language Models," arXiv, 2024.

[6] J. Kim, Y. Lee, Y. Han, S. Jung, and H. Choi, "Does Incomplete Syntax Influence Korean Language Model? Focusing on Word Order and Case Markers," *arXiv*, 2024.

[7] J. Yang, J. Kim, C. Seo, D. Hwang, K. Kim, and R. Kim, "Automatic Textual Data Transformation for Enhancing F1-score on Classification," in Proc. *International Symposium on Advanced and Applied Convergence*, vol. 24, pp. 2-6, 2024.

[8] S. Park et al., "KLUE: Korean Language Understanding Evaluation," arXiv, 2021.

[9] KLUE-benchmark, "KLUE: Korean Language Understanding Evaluation," Github Repository, Accessed Nov. 20, 2024, https://github.com/KLUE-benchmark/KLUE.

[10] S. Rigatti, "Random forest," J. Insurance Med., vol. 47, no. 1, pp. 31-39, 2017.