Improvement of classification task using data augmentation and LLM(in progress)

Yeonsoo Kim
Aiffel/NLP track
yeonsookim824@gmail.com

Abstract

With the advancement of natural language processing technology, it has become possible to perform a variety of tasks. However, new English-based technologies and models are typically released rapidly, making it challenging to apply them directly to Korean. Nevertheless, Korean-specific data processing techniques and models are also being developed at a fast pace, making it feasible to apply certain natural language processing tasks to Korean. Among these, this paper describes a project focused on building a classification task model and improving its performance. In this project, the team utilizes the Dataset of Korean Threatening Conversation (DKTC) data. They train the model using approximately 4,000 training data samples and evaluate its performance with 400 test data samples. Since the goal of this project is to enhance performance, research is conducted on data preprocessing techniques and the models used. The team adds stopword removal to basic data preprocessing techniques and applies two data augmentation methods: KorEDA and Back Translation. To achieve good performance, they use Language Model (LLM) models and experiment with three models that are known to perform well with Korean data: klue/bert-base, skt/kogpt2-base-v2, and KcELECTRA-small. The experiments involve making changes such as additional preprocessing (stopword removal), data augmentation, and model replacement to compare the results. Ultimately, the klue/bert-base model with stopword removal achieves the highest accuracy at 0.92. Throughout the project, the team gains an understanding of the importance of both the quantity and quality of the data used. Additionally, they recognize the need for further research to adapt new natural language processing techniques like EDA and LLM to the Korean language.

1 Introduction

Natural language processing technology is advancing rapidly and is increasingly performing a variety of tasks. By collecting large amounts of natural language data, preprocessing it, and training models, efforts are being made to improve performance. However, most rapidly emerging natural language processing technologies are developed based on English, which makes it challenging to apply them to languages like Korean, which have entirely different language structures. As a result, there are ongoing efforts to develop Korean versions of these technologies and models, but the limitations arising from language structure differences cannot be ignored.

The task addressed in this paper is the classification of Korean language data. The ultimate goal is to create a deep learning model that classifies four types of conversations using the DKTC (Dataset of Korean Threatening Conversation) training data: threats, extortion, workplace harassment, and other harassment. In this project, Korean-specific augmentation techniques and pretrained models were used.

Preprint. Under review.

The DKTC training data used for training is in a conversational format. Initially, performance is evaluated using basic data preprocessing and models. Subsequently, various preprocessing techniques and LLM (Language Model) models are employed to attempt performance improvement.

2 Method

There can be various ways to improve the performance of classification task. It can be improved by adding new dataset, training with various models, or changing parameters. Our team has decided to adapt these all 3 methods to improve our classification task. We basically chose a baseline model as **klue-bert**, and tried to experiment with it. After experimenting with baseline model, we adapted those methods to the two other new models.

2.1 data preprocessing

Data preprocessing must be the most fundamental, but crucial step. At first, we used only basic processing method. But later, we tried preprocessing without stopwords again and it led to a meaningful task improvement. Here's the detail how we did data preprocessing.

2.2 models

As our goal was explicit, getting the best performance, we concluded using representative pretrained models. So we searched some korean-based pretrained models and compared their accuracy. As a result, klue-bert, skt/kogpt2-base-v2, KcELECTRA-small were chosen as our experiment models.

- klue/bert-base:
- skt/kogpt2-base-v2:
- KcELECTRA-small:

2.3 Data Augmentation

We expected that Data Augmentation method would be the most influential method regarding to task improvement. So, we tried two augmentation techniques: 1) KorEDA and 2) adding new dataset+back translation.

1) KorEDA
KorEDA is a korean version of EDA(Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks), changing wordnet part as korean(Korean WordNet(KWN)). EDA processes data in four ways: sr(synonym replacement), ri(random insertion), rs(random swap) and rd(random deletion). When we first adapted KorEDA to ML model(Naive Bayes model), it seemed like the accuracy was improving. But with pretrained models, it showed no improvement.

2) new dataset + back translation
Analyzing resource dataset, we found out that the dataset itself is too small. We found a dialogue dataset from AIHUB and processed it into the form of our resource data and combined them. Plus, we did back translation. Back translation is a method used to generate a synthetic dataset by translating textual information into another language and translating it back into the original language. Recognizing that the essence of the sentence remains intact even after undergoing the process of back translation, we concluded that it could be applied to our task.

3 Result

The results of experiment are as follows:

"Accuracy (a)" is the result obtained by applying only basic preprocessing to the original data with a batch size of 16, and "Accuracy (b)" is the result obtained by applying augmented data with a batch size of 16. Up to this stage, the klue/bert-base model showed the best performance, but it couldn't surpass a certain level of improvement. Since we had conducted data preprocessing at a very basic level, we decided to add stopword removal as an additional step. The result of

Table 1: Result

Model	Accuracy(a)	Accuracy(b)	Accuracy(c)
KcELECTRA-small	0.815	0.7975	-
skt/kogpt2-base-v2	0.8625	0.8775	-
klue/bert-base	0.90	0.905	0.92

removing stopwords and increasing the batch size to 64 is represented as "Accuracy (c): 0.92." This was not only the highest achievement within our team's experiments but also among all the teams participating in the project. Regarding data augmentation techniques, back translation had a significant impact on performance improvement. There was an expectation that EDA techniques, which are known to be highly beneficial for data augmentation in English-based text, would yield noticeable results, but this was not the case. This may be because the four techniques (rs, ri, rd, sr) used for data augmentation in English text did not apply well to Korean data. In fact, when each technique was separately applied to Korean data, it was observed that it often introduced a significant amount of noise. Importing additional data from AIHub and performing back translation proved to be an effective data augmentation technique. It is worth noting that when importing data from external sources, the format and content should ideally be similar to the existing data. As long as this consideration is kept in mind, data augmentation can be quite helpful. Back translation, which was applied for the first time in this project, appears to be a useful augmentation technique. It not only increases the quantity of data but also maintains semantic consistency, making it a valuable augmentation technique both quantitatively and semantically.

4 Conclusion

The purpose of this paper was to construct a model for performing a classification task using the DKTC dataset and to improve its performance. Data preprocessing and augmentation were applied to fit the Korean language, and models trained in Korean were utilized. It was observed that both data preprocessing techniques and the choice of model significantly influenced the model's performance. For data augmentation, two methods, KorEDA and Back Translation, were employed, with the latter method yielding better results. In the case of KorEDA, it was an attempt to adapt English-based EDA techniques, with only Wordnet transformed into a Korean version, which did not yield the expected results. This might be attributed to the fact that English-based EDA methods, which involve sr (synonym replacement), ri (random insertion), rs (random swap), and rd (random deletion), do not induce significant changes in meaning in Korean sentences. While EDA techniques are quite effective for data augmentation in English text, it appears that further research is needed for Korean EDA. Back Translation, on the other hand, was a technique that significantly improved performance.

One slightly disappointing aspect of the project was the limited amount of data available, with only around 4,000 training data samples and 400 test data samples. Additionally, during the early stages of the project, the team engaged in labeling to expand the dataset, providing an opportunity to closely examine the data. It was observed that many instances of ambiguously or incorrectly labeled data were mixed in.

References

[1] Chen, J. &Tam, D., Raffel, C. Bansal, M., & Yang, D. (2023). An empirical survey of data augmentation for limited data learning in NLP. Transactions of the Association for Computational Linguistics, 11, 191-211.