The 6th International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2022)
December 26-28, 2022. Pattaya, Thailand.

Masked language modeling-based Korean Data Augmentation Techniques Using Label Correction

Myunghoon Kang[1], Jungseob Lee[1], SeungJun Lee[1], Hyeonseok Moon[1],
Chanjun Park[2], Aram So[3*]

[1] Korea University, Computer Science
[2] Upstage
[3] Human-inspired AI Research
{chaos8527, omanma1928, dzzy6506, glee889, aram}@korea.ac.kr
chanjun.park@upstage.ai

Abstract. Recently, many researchers have focused on utilizing masked language modeling (MLM) as a data augmentation tool. However, previous works didn't consider the error propagation problem caused by the MLM-based data augmentation approach, randomly replacing crucial tokens in the sentence. To mitigate this problem, this paper proposes a re-labeling module capable of correcting the label of the augmented data generated by MLM-based Korean data augmentation techniques. We found that the proposed model outperforms the existing methods in the KLUE-STS task with a smaller label-corrected augmented dataset.

Keywords: Data Augmentation, Masked Language Model, Semantic Textual Similarity, Deep Learning

1 Introduction

Data augmentation is a technique that increases the amount and variety of the dataset without explicitly collecting more data[1]. Previous works[2,3] utilized Masked Langauge Model (MLM) for text data augmentation. However, previous works didn't consider the label aligning with the augmented data. This leads to an error propagation problem, especially in the case of the paired input dataset, that model learns supervised error cases and ends up with poor prediction performance.

In this study, we proposed a re-labeling module, which corrects the label of the augmented data generated by MLM-based augmentation methods. We constructed a re-labeling module with a re-labeling model and gate function. Re-labeling model is the KLUE-bert-base[1] model fine-tuned on the KLUE-NLI[4] dataset. Once the augmented data is fed to the re-labeling model, the re-labeling model predicts the label of the input. Then, the gate function filters the uncertain corrected cases based on the confidence threshold *t*.

[1] https://huggingface.co/klue/bert-base
* Corresponding Author

We conducted a comparative experiment to evaluate the effectiveness and validity of the proposed re-labeling module. For the experiment, we constructed three cases for measuring the effect of the data augmentation methods, based on the baseline, MLM augmented, and the re-labeling module. We use the KLUE-STS[4] dataset and measure the performance using the micro F1 score. It is shown that our proposed re-labeling module outperforms MLM augmented case with a smaller label-corrected augmented dataset.

2 Experimental Results

We use the train set (11,668) of the KLUE-STS for training and data augmentation. For data augmentation, we randomly masked the token with a 15% probability. If the token is replaced with [MASK] token, we augmented 5 sentences using the predicted tokens from the KLUE-bert-base model. For the re-labeling module case, we fed augmented datasets to the re-labeling module for label correction with a confidence threshold of t=0.9. For the evaluation, we use the dev set (519) and evaluated binary classification results with a micro f1 score.

The experimental results are shown in Table 1. As a result of this experiment, the re-labeling module showed better performance than the MLM augmented and baseline. Despite the small data size of re-labeling, our proposed method, the re-labeling module shows great performance compared to MLM augmented with larger data size. This shows the effectiveness of our proposed method.

Table 1. Comparison between baseline, MLM augmented and re-labeling module on STS task, evaluated with mico F1 score.

Model	Train dataset size	Dev micro F1
Baseline	11,668	0.818
MLM augmented	51,983	0.822
Re-labeling module	16,191	**0.828**

3 Conclusion

In this paper, the effectiveness and validity of the re-labeling module were verified by applying it to the KLUE-STS task. In the future, we plan to advance MLM based approach for data augmentation which awares the core semantic tokens in the sentence.

4 Acknowledgement

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 research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)"

References

1. Feng, Steven Y., et al. "A Survey of Data Augmentation Approaches for NLP." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021.
2. Kim, Kihun, et al. "BERT-based Data Augmentation Techniques for Korean Coreference Resolution." *Annual Conference on Human and Language Technology*. Human and Language Technology, 2020.
3. Kim, Jungwook, et al. "Named Entity Recognition based on ELECTRA with Dictionary Features and Dynamic Masking." *Annual Conference on Human and Language Technology*. Human and Language Technology, 2021.
4. Park, Sungjoon, et al. "KLUE: Korean Language Understanding Evaluation." *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*. Advances in Neural Information Processing Systems, 2021.