The 7th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2023)

Improving TOEIC Problems Solving Model Performance through Data Augmentation using WordNet

Jeongwoo Lee[1], Aiyanyo Imatitikua Danielle[2], Yeongwook Yang[3*]
[1]Dept of Computer Science and Engineering, Korea University
[2]Human-inspired AI Research
[3]Division of Computer Engineering, Hanshin University
{time79779, titi}@korea.ac.kr, yeongwook.yang@gmail.com

Abstract

Recently, a lot of research has been conducted on understanding text and inferring answers, and a representative example is research on machine reading. Although various datasets have been released related to machine reading comprehension, there are almost no officially released datasets for TOEIC, which has been widely used from the past to the present to evaluate people's English proficiency, and research on this is also actively underway. not. Accordingly, in this study, we would like to propose a data augmentation technique to improve the performance of machine reading models in the current data-scarce situation. The proposed method uses WordNet to augment data very simply and efficiently similar to actual TOEIC problems based on synonyms and antonyms, and the significance of the method was confirmed through experiments. Through this study, we aim to solve the problem of insufficient data for TOEIC and achieve excellent human-level performance.

Keywords: Deep Learning, Natural Language Processing, Machine Reading Comprehension, Data Augmentation

1. Introduction

Currently, deep learning technology has developed greatly and is achieving great results in various fields, and there are many models that surpass human levels. However, not much research is being done to solve actual tests that evaluate human abilities with deep learning models. In particular, TOEIC has been widely used as a standard for evaluating a person's English ability from the past to the present, but deep learning models are not showing great results in this regard.

We believe that the reason deep learning models do not produce good results is a data problem. Currently, there is a lot of data for solving human problems, but there is almost no publicly available data for machine learning. Accordingly, we proposed a training data augmentation technique for the model to improve the deep learning model's ability to identify relationships between words in sentences and improve the performance of the TOEIC problem-solving model, and analyzed the results of using this method.

Among the various parts of TOEIC, we conduct research on Part 5, a cloze test [1] that can be meaningfully used to evaluate language skills [2]. In addition, an officially open dataset is essential to objectively verify the performance of the model. While there are currently no official datasets for Part 5, the dataset for Part 5 exists on Kaggle. It is suitable for objective performance verification[1]. However, since there are only 3,625 pieces of Kaggle data, it cannot be said to be a sufficient amount for machine learning. Therefore, in this paper, we propose a very simple yet efficient data augmentation technique based on synonyms and antonyms using WordNet, which allows us to create an excellent deep learning model in the TOEIC problem solving task.

2. Proposed Method

TOEIC's Part 5 short-sentence fill-in-the-blank question consists of a sentence containing a blank, a correct answer option corresponding to the blank, and an incorrect answer option. It is important to appropriately create incorrect answer options as they may cause confusion when choosing the correct answer. In this study, inspired by these characteristics, we propose a method to augment data based on synonyms and antonyms.

TOEIC's Part 5 Short Sentence Blank Filling Problem is a problem in which one sentence with a blank space is

*Corresponding author.

[1]https://www.kaggle.com/tientd95/toeic-test

Table 1. Experimental results on Kaggle TOEIC data and Synonyms/Antonyms-based augmentation data

Data	Accuracy
Kaggle TOEIC data	79.06%
Synonyms/Antonyms-based aug data + Kaggle TOEIC data	86.77%

given, and the word, phrase, or clause that best fits the blank space is selected from among four options. Similarly, to augment the data, we use English sentence data from AI Hub's Korean-English translation (parallel) corpus to generate data by cutting the sentence data into word units and replacing one of them with a blank space. At this time, WordNet is used to extract synonyms and antonyms for the correct answer options to be entered in each blank, and three are randomly extracted from the extracted set of synonyms/antonyms to form incorrect answer options. Sometimes synonyms and antonyms are the correct answer, but in this study, in order to maximize the simplicity and efficiency of data augmentation, the correct answer in the original sentence is set as the most correct answer, and when such data becomes countless, the deep learning model's sentence I assumed that my ability to understand relationships between words would improve, and I proved this through experiments.

3. Experiments

The model used in this study was BERT [3], and learning was conducted using bert-large-uncased. The batch size was set to 128, max epoch was set to 50, and learning rate was set to 3e-5, and the GPU used was NVIDIA RTX A6000.

Learning was conducted on Kaggle TOEIC data and synonym/antonym-based augmented data generated using the method proposed in this paper. The test was conducted with 363 TOEIC data from Kaggle, and Table 1 shows the results of the experiment.

As can be seen from the results of the experiment, it can be seen that the synonym/antonym-based augmented data generated by the method proposed in this paper has a significant effect on performance improvement. Finally, when the experimental results are confirmed, it can be confirmed that it performs much better to generate data by the method proposed in this paper and proceed with learning first and learn additional Kaggle TOEIC data than to proceed with learning only with Kaggle TOEIC data.

4. Conclusions

In this study, we conducted research on ways to improve the performance of machine reading models in situations where there is insufficient data on TOEIC, which is used to evaluate people's English proficiency, and proposed a synonym/antonym-based data augmentation technique. Sometimes synonyms and antonyms are the correct answer, but in this study, we focused more on maximizing the simplicity and efficiency of data augmentation. We set that the correct answer in the original sentence is the most correct answer, and when such data becomes numerous, deep learning is used. Through experiments, it was confirmed that the model's ability to understand relationships between words in sentences improved. Through this, even if there is a problem of insufficient data, the performance of the machine reading model can be improved and excellent performance can be obtained.

Acknowledgements

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). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C2004868).

Reference

[1] W. L. Taylor, ""cloze procedure": A new tool for measuring readability," *Journalism quarterly*, Vol. 30, No. 4, pp. 415–433, 1953.

[2] J. Jonz, "Cloze item types and second language comprehension," *Language testing*, Vol. 8, No. 1, pp. 1–22, 1991.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.