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

Data Augmentation Schemes For Machine Reading Comprehension

Jeongwoo Lee[1], Aiyanyo Imatitikua Danielle[2*1]

[1]Dept of Computer Science and Engineering, Korea University
[2]Human-inspired AI Research
{time79779, titi}@korea.ac.kr

Abstract. Recently, many studies have been conducted to infer answers based on contents. A typical example is machine reading comprehension research, and several related datasets are also publicly available. However, few datasets exist for tests that evaluate people's reading comprehension skills, such as Korean College Scholastic Ability Test or TOEIC. For this reason, in tests evaluating human reading comprehension ability, research to solve it with an artificial intelligence model is not actively conducted, and the performance improvement of the model's reading comprehension ability is limited. Accordingly, we try to solve the problem of insufficient data set for reading comprehension by proposing 3 data augmentation techniques.

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

1 Introduction

In this paper, 3 data augmentation methods are proposed to solve the problem of the data scarcity for tests evaluating human reading comprehension. These are methods using Round Trip Translation[1], Word Embedding[2, 3], and Wordnet[4], respectively.

The reading comprehension problem dealt with in this paper refers to a problem that includes several contents such as additional passages and references, as well as question and options, such as the Korean section of the Korean College Scholastic Ability Test.

* Corresponding author.

2 Proposed Method

2.1 Data Augmentation Using Round Trip Translation

The first data augmentation method proposed in this paper is a data augmentation method using Round Trip Translation. This refers to a method of generating new data having the same meaning as the original sentence but having the different sentence structure by extracting the sentence from the reading comprehension problem data, translating the sentence into another language, and translating the result back to the original language. Through this, the data can be expanded by augmenting the sentence data existing in the question, the options, and several contents.

2.2 Data Augmentation Using Word Embedding

Second, we propose the data augmentation technique using Word Embedding. This is a method of constructing new data by measuring the embedding similarity between words and replacing the word in the sentence with the word having the similar embedding value.

The specific method of data augmentation using Word Embedding is as follows. The sentence is extracted from the question, options and several contents of the reading comprehension data, and nouns to be replaced with synonyms are extracted by applying Part of Speech (POS) tagging using KoNLPy's[5] Mecab-ko morpheme analyzer in the corresponding sentence. And, using FastText and Word2vec, the similarity between all defined words and the nouns extracted before is measured, and the words with the highest similarity are extracted. By randomly sampling some of the similar words extracted in this way and replacing the words in the existing sentence to generate variant sentences in various combinations, sentences with similar meanings to the existing sentences can be obtained.

Through this, the sentences of the question and several contents can be transformed into other sentences with similar meanings, and the correct and incorrect answers of the options can be augmented by transforming them into other options with similar meanings, respectively.

2.3 Data Augmentation Using Wordnet

In this method, we propose a method of finding synonyms and antonyms of words using Wordnet, and augmenting data based on them. The method for augmenting data using synonyms obtained via Wordnet is similar to Section 2.3. Therefore, in this method, a method of augmenting data using antonyms will be additionally dealt with.

The process of augmenting data with antonyms obtained using Wordnet is as follows. From the correct answer options of the reading comprehension data, the noun to be replaced with the antonym is extracted through POS tagging using KoNLPy's Mecab-ko morpheme analyzer, and the antonym of the extracted noun is obtained

using Wordnet. By randomly sampling some of the antonyms obtained in this way, by replacing the words of the existing correct answer options to generate variant options in various combinations, incorrect answer options with different meanings from the existing correct answer options can be obtained.

Through this, by generating not only the given options in the multiple-choice reading comprehension problem but also more options, learning can proceed with more diverse combinations of options, and thus the performance of the model can be expected to be improved.

3 Conclusion

In this study, 3 data augmentation techniques using Round Trip Translation, Word Embedding, and Wordnet were proposed to solve the problem of insufficient datasets for reading comprehension problems. Through the methods proposed in this study, it is expected that the model can obtain excellent performance through sufficient training data. In the future, we plan to conduct experiments on various machine reading comprehension tasks by using the data augmented through the methodologies proposed in this study.

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).

References

1. Somers, H. (2005, December). Round-trip translation: What is it good for?. In Proceedings of the Australasian Language Technology Workshop 2005 (pp. 127-133).
2. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.
3. Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504.
4. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. International journal of lexicography, 3(4), 235-244.
5. Park, E. L., & Cho, S. (2014). KoNLPy: Korean natural language processing in Python. In Annual Conference on Human and Language Technology (pp. 133-136). Human and Language Technology.