데이터사이언스학 석사 학위논문

Shaking Attention Scores in Pretrained Transformers

트랜스포머의 어텐션 스코어 조작에 관한 연구

2023년 2월

서울대학교 대학원

데이터사이언스학과 데이터사이언스학 전공

김 종 원

Shaking Attention Scores in Pretrained Transformers

트랜스포머의 어텐션 스코어 조작에 관한 연구

지도교수 이 재 진

이 논문을 데이터사이언스학 석사 학위논문으로 제출함

2022년 12월

서울대학교 대학원

데이터사이언스학과 데이터사이언스학 전공

김 종 원

김종원의 석사 학위논문을 인준함

2023년 1월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

Abstract

Although Korean has distinctly different features from English, attempts to find a new Transformer model that more closely matches Korean by reflecting them are insufficient. Among the characteristics of the Korean language, we pay special attention to the role of postpositions. Agglutinative languages have more freedom in word order than inflectional languages, such as English, thanks to the postpositions. This study is based on the hypothesis that the current Transformer is challenging to learn the postpositions sufficiently, which play a significant role in agglutinative languages such as Korean. In Korean, the postpositions are paired with the substantives, so paying more attention to the corresponding substantives seems reasonable compared to other tokens in the sentence. However, the current Transformer learning algorithm has many limitations in doing so. Accordingly, it is shown that the performance of the natural language understanding (NLU) task can be improved by deliberatively changing the attention scores between the postpositions and the substantives. In addition, it is hoped that this study will stimulate the research on new learning methods that reflect the characteristics of Korean.

Keyword : Transformer, attention score, natural language processing, NLU, agglutinative language, substantives, postposition,

Student Number : 2021-24334

Table of Contents

Chapter 1. Introduction

The influence of Transformer (Vaswani et al., 2017[1]) in the field of natural language processing (NLP) is excellent. The first Transformer evolved into many derivative models and achieved state-of-the-art (SOTA) in many downstream tasks. Several years have passed since the Transformer was proposed, but newly derived models are still being offered, and it is not easy to imagine NLP without the Transformer.

Researches on Transformers mainly propose new models or training methods to solve specific problems (Yang et al., 2019[2]; Conneau et al., 2019[3]; Joshi et al., 2020[4]). Recently, attempts to apply them to various fields, such as images beyond the boundaries of NLP, have been active (Dosovitskiy et al., 2020[5]; Ramesh et al., 2021[6]). In comparison, analysis of the internal structure of Transformers, especially on self-attention, which is a core structure, needs more research (Lin et al., 2019[7]; Serrano and Smith, 2019[8]). All the more so considering that it has been a significant quantity of time since the Transformer was proposed. There needs to be more than the latest research findings to explain the fundamental reasons Transformers achieve SOTA in various tasks. The Transformer only relies on the outcome to produce better-derived models without a clear understanding of the causal relationship between the attention scores and the final performance.

In addition, most of Transformer research is centered on Western languages such as English. In Korea, the size of researchers and the study on correspondence between Transformer and Korean is poorer. Korean is an agglutinative language in which affixes such as postpositions play an essential role. Although the positions of the subject, object, and adverb have been arbitrarily changed in the

sentences of "나는 너를 학교에서 보았다 (I saw you at school)", "너를 나는 학교에서 보았다", "나는 학교에서 너를 보았다", "학교에서 나는 너를 보았다", "너를 학교에서 나는 보았다", each one is grammatically OK and has the same meaning. The position of the subject and object can be exchanged due to the postpositions such as "~는" and "~를". It is a significant characteristic of agglutinative languages. However, in inflectional languages like English, changing word order within a sentence in this way is much more limited. The word itself is transformed or its role changes depending on its position in the sentence.

Self-attention, the core structure of Transformer, calculates the individual correlation with all words (key) including itself for each word (query) in the input sentence as the attention scores. All words in a sentence are calculated at once with query and key matrix, which is the direction the Transformer proposed to improve speed through parallel processing aimed at in the first place. However, whereas recurrent neural networks (RNN) such as LSTM (Long Short-Term Memory, Hochreiter and Schmidhuber, 1997[9]) process words one at a time and naturally reflect word order information, self-attention's parallel processing does not match the general characteristics of language where word order is essential. So Transformer requires absolute word position (Vaswani et al., 2017[1]; Devlin et al., 2018[10]; Huang et al., 2020[11]) or relative position information between words (Shaw et al., 2018[12]; Dai et al., 2019[13]; Raffel et al., 2020[14]), which is called positional encoding (PE).

Word order within a sentence is essential to understanding language (Sutskever et al., 2014[15]). However, several PE algorithms proposed so far have progressed in a direction that matches well with inflectional languages such as English. It is because the achievement of SOTA in the English-centric corpus has a decisive impact on the model's reputation. Earlier, it was said that inflectional languages should be more strictly

followed in word order compared to agglutinative ones such as Korean. However, whether the PE information developed like this will also perform optimally in Korean is doubtful. When dealing with English and Korean, there may be elements that need to be treated differently to reflect the different language characteristics. However, related research has not been conducted so far. Since this topic is not so attractive to English-speaking researchers, it may be an assignment only for Korean researchers. However, it is a regrettable reality that it is not easy even to follow the achievements made in the English-speaking world.

This study began with whether a higher performance could be obtained if the characteristics of Korean as an agglutinative language were reflected in Transformer, which has dramatically developed in the inflectional language environment. The Korean language characteristic that we pay attention to here is the role of postpositions. As mentioned earlier, postpositions in Korean play a considerable role in a sentence. Thanks to postpositions, original meaning can be maintained even if a word's position is changed. If so, the postpositions in the Transformer should pay more attention to the corresponding substantives than other tokens. Unfortunately, however, there is no way to verify whether the attention of the postposition has been properly learned with the research achievements so far. It is just known that attention heads have various viewpoints (Clark et al., 2019[16]), which means that the attention score between specific tokens may be inconsistent across multiple heads. In other words, it is difficult to tell whether the relationship between postpositions and substantives has been sufficiently learned with the simple metric such as frequency of the attention score pattern.

Although it is difficult to verify whether the postposition learned substantives properly, it is clear that the Korean language does not

match well with the current Transformer learning algorithm, which will be explained in Chapter 3. What will happen to NLU task performance if we deliberatively shake the attention scores that have not been sufficiently learned? This study observes NLU performance after manipulating the attention scores between substantives and postpositions and between prefix/adnominals and substantives. Various experiments are performed, such as manipulating the attention scores while finetuning or training from scratch. In addition, a test is also conducted for the case where the attention score is manipulated for arbitrary tokens rather than words with substantives and postpositions. These tests confirm that NLU performance can be improved by intentionally manipulating attention scores. This study makes the following contributions.

- It shows that the performance of the NLU task can be improved by directly manipulating the Transformer's attention scores.
- The current Transformer training algorithm needs to be revised to reflect the characteristics of Korean, which have a significant role in postposition.

Chapter 2. Related work

Clark et al. 2019[16] conducted a more systematic study of the attention weight aspects of BERT (Devlin et al., 2018[10]). Through this, several common patterns among attention heads were found. It is about the position of key token that query token mainly pays attention to. For example, it pays more attention to the token immediately following the current query token or the [SEP] token at the end of the sentence. In particular, it is argued that the case where the attention weight of the [SEP] token has a large value corresponds to a no-op. In addition, it is shown that there is an aspect in which one head captures one specific relationship well rather than capturing several ones.

Kovaleva et al. 2019[17] studied better capturing the linguistic characteristics by checking the self-attention pattern for each head of the BERT model. They also argue that attention tends to be overparameterized. Hao et al. 2018[18] proposed a new metric that gives a larger value as the effect on the final output increases, independently of the attention weight. Attention weight, like [SEP], does not affect the output but has a large value, so the need for a different metric was argued. It is also shown that the attention head considered more critical can be found. However, it does not include specific details about the correlation between semantically or grammatically close tokens within a sentence.

Pruthi et al. 2020[19] researched manipulating attention weights. It is a method of pre-determining disallowed words and adding a penalty term to the objective function. It is not directly manipulating the attention weights but manipulation in that the attention weights that should

originally come out based on the input data are changed due to the penalty term. Also, since it is to analyze the model's performance using only the remaining allowable words, it is different from what this study intends. Li et al., 2018[20] devised a metric to quantitatively confirm that each attention head captures different characteristics of a sentence. Jawahar et al. 2019[21] conducted a study on whether BERT can learn a language, especially the linguistic characteristics of English. It was shown that there is a difference depending on the attention layer, and semantic characteristics were mainly found in the upper layer.

Chapter 3. Korean and Transformer

As a result of the current Transformer study, there is no general way to analytically check whether the postposition, which plays a crucial role in Korean, properly pays attention to the corresponding substantives. However, considering the characteristics of Korean and the current Transformer learning algorithm, it can be inferred that certain limitations exist. As mentioned earlier, the token's position (PE) in a sentence is essential for learning. As for position information, even if the absolute position is used, the relative distance information between two tokens plays a vital role (Vaswani et al., 2017[1]). Korean can easily move the word's position thanks to the postposition as in the previous example of "나는 너를 학교에서 보았다 (I saw you at school)". It is because the postposition retains the part of speech (POS) of the word even if its position is changed.

On the other hand, in English, POS is determined by the position of the word in the sentence ("Tom likes Jane" vs. "Jane likes Tom"), or the word itself changes when the POS is changed ("I like him" vs. "He likes me"). Therefore, relative distance information between words is much more important in English than in Korean. These characteristics of English are reflected in the current PE algorithm. However, of course, it does not cover the characteristics of Korean, which are more unrestricted in the movement of word positions.

In addition, the position of "나는" can be easily moved, whereas the relative position of "나" and "~는" must be fixed. That is, the two traits coexist, whether the tokens' relative position is essential or not. PE does modeling a fundamental property of language: word order. The current

PE, which treats the relative distance information between each token with the same importance for all tokens, does not correctly reflect the characteristics of Korean, which has both aspects. Considering these points, we can not say that Korean can be sufficiently learned with the current Transformer learning algorithm.

The fact that the word with substantives and postpositions can move more freely within a sentence does not match well with masked language modeling (MLM), which is the core learning algorithm of the Transformer encoder model as well as PE. However, as in English, the characteristic that words change depending on POS or their relative position is fixed to some extent helps learn through MLM to predict masked tokens using surrounding words.

However, in Korean, since the postposition itself plays a prominent role in the meaning of a sentence, it is challenging to predict the masked token just with the given surrounding words. It is because the postposition defines its role of itself. For example, if "~는" and "~를" are masked in "나는 너를 학교에서 보았다 (I saw you at school)", how can we accurately infer these two tokens with only the surrounding words? No matter how much data is learned, it does not seem enough. Because "나를 너는 학교에서 보았다 (You saw me at school)" is also a correct sentence. As such, if there are many correct sentences, the learning effect through the corpus is inevitably lower than expected. Therefore, a new learning algorithm is required to cover the postposition's role in addition to the existing MLM.

Chapter 4. Methodology

Shaking attention scores

There are two ways to shake the attention scores in this study. The first is like purposely changing the attention score of "나" when "~는" is a query as in "나는 너를 학교에서 보았다 (I saw you at school)". The second is randomly selecting tokens in the sentence to change the attention scores. Below, we will mainly explain the former, which is changing the attention scores for the case of substantives and postpositions. The models that performed the test were imported from Hugging Face (HF)[①] or trained from scratch.

Attention weights are obtained by passing attention scores through softmax to form probabilities. If a specific attention weight is directly manipulated, correcting the attention weight of other tokens must be followed. So, for convenience, attention scores were manipulated instead of attention weights. Manipulations include both raising and lowering the original attention scores. Since the attention scores can be negative, the absolute value of the attention scores to be manipulated is multiplied by a positive or negative variable; we call it boost_factor (bf), and multiplied result is added to the original attention scores. If the sign of boost_factor is positive, it is larger than the original attention score, and if it is negative, it is smaller. Expressed in code, it is equivalent to Equation (1). b_mtrx has the same size as the attention_scores matrix, and all other positions except for the manipulated positions are 0. For example, the attention scores will be manipulated at the positions of key

① https://huggingface.co/models?language=ko&sort=downloads

"나 (I)" for query "~는" and key "너 (You)" for query "~를" for the sentence "나는 너를 학교에서 보았다 (I saw you at school)".

attn_score += | attn_score ⊙ b_mtrx | * bf (Eq. 1)

The tokenizer of the HF model cannot get POS, such as postpositions. Therefore, an additional tokenizer is used to find POS. After setting all elements of b_mtrx to 0 as the initial value, if the obtained POS is a postposition and the tokens generated by the two tokenizers are the same, a value other than 0 is written in the corresponding position of b_mtrx. Suppose different values are written according to the type of postposition. Then, the change of attention scores can be set differently according to the postpositions, even under the same boost_factor.

There are several combinations regarding the sign of the boost_factor that manipulates the attention scores in the pretrained model. For example, apply a negative to finetuning and a positive to inference. A total of four combinations were tested for each downstream task, and the combination with the best performance was applied. In finetuning, negative and positive, and in inference, 0 or positive are tested in combination. If 0 is used, the attention scores do not change.

Hyperparameters

The postpositions can be classified into nine subcategories, for example, the nominative (JKS) and the objective (JKO). Among the postpositions, tests such as a single application or a single exclusion were performed for each postposition to find a more critical postposition. In addition, various tests were performed with different combinations. Finally, the postpositions were divided into two groups according to test results. For

obtaining the best performance on average, JKS (nominative case), JKO (objective case), and JX (auxiliary case) are group 1. The remaining six postpositions are tagged as group 2. Prefixes and adnominals are also tested as group 3 to change the attention scores. In the end, the boost_factor of Eq. 1 is applied to groups 2 and 3, and the boost_factor*boost_prem is applied to group 1 instead of boost_factor. boost_prem is the additional hyperparameter.

After finding the optimal value by applying general hyperparameters such as learning rate, boost_factor and boost_prem are used to obtain additional performance. When manipulating the attention scores between substantives and postpositions, boost_factor and boost_prem are hyperparameters. If it is the manipulation of attention scores for randomly selected tokens, only boost_factor is used as a hyperparameter.

Self-attention layer test

Several tests were performed to determine the optimal layer for manipulating the attention scores. First, for all layers, the test was performed for only one layer or excluding only one layer, respectively. In addition, another test was also run in which attention score changes were added sequentially, starting from the top layer in a downward direction or, conversely, from the bottom layer upwards. Lastly, some test was performed to increase or decrease the boost_factor for all layers linearly. In conclusion, applying the same boost_factor to all layers gave the overall best results, and the test results presented in this study were executed under this condition.

Test model and downstream task

Downstream tasks were performed on the four pretrained models from HF and the six models pretrained from scratch. All models were trained with the Korean corpus. The HF models are 1) klue/roberta-base trained on the RoBERTa base model (Liu et al., 2019[22]), 2) klue/bert-base trained on the BERT base (Devlin et al., 2018[10]), 3) klue/roberta-small trained on the RoBERTa small, 4) monologg/koelectra-base-v3-discriminator trained on ELECTRA base (Clark et al., 2020[23]). All the models pretrained from scratch were based on RoBERTa small. Unlike the base model with 12 layers and 12 heads, the RoBERTa small has 6 layers and 12 heads. The dataset used for pretraining is '신문 말뭉치 (newspaper corpus)' and '문어 말뭉치 (written corpus)' in '모두의 말뭉치' from National Institute of Korean Language②. It has a total capacity of 16.2 GB and 3.2 billion tokens.

There are seven downstream tasks. Three of them are from the 2021 AI Language Proficiency Contest (National Institute of Korean Language③): 1) Distinguishing homographs (WiC, Word In Context), 2) Inference of causal relationship (CoPA, Choice of Plausible Alternatives), 3) Decision questions (BoolQ, Boolean Questions). The rest of them are from Korean Language Understanding Evaluation (KLUE, Park et al., 2021[24]): 4) Topic Classification (TC) or YNAT (Younhap News Agency news headlines for Topic Classification), 5) Semantic Textual Similarity (STS), 6) Natural language inference (NLI), and 7) Machine Reading Comprehension (MRC).

These downstream tasks do not disclose the test set or ground truth.

② https://corpus.korean.go.kr/request/corpusRegist.do

③ https://corpus.korean.go.kr/task/taskDownload.do?taskId=1&clCd=END_TASK&subMenuId=sub02

Therefore the evaluation set is divided in half and used as the test set so that the result may differ from the actual test set. Nevertheless, for KLUE tasks, The finetuning test was performed at about the best performance level, which was tested and provided for reference by KLUE. What is important is what extent additional performance will be when the attention scores are shaken even under the close state to the maximum performance.

Pretraining from scratch

All six pretrained models have the same conditions except for the training method.

1) A model generally learned without manipulating attention scores. It is used as a baseline for performance comparison with other models.

2) Learning was performed by setting the boost_factor that changes the attention scores between substantives and postpositions to 0.3. In this case, it has larger values than the original attention scores.

3) It is the same as 2) except for boost_factor to -0.3. It will have smaller attention scores than the original.

4) boost_factor 0.2 was applied to random select tokens with a probability of 8%.

5) The positional encoding of the postposition token was set to the same as substantives'.

6) Two tokens of the substantives and the postposition were replaced with a new token by summing these two embeddings.

Earlier, it was explained that the current Transformer learning algorithm could not be considered to learn the postpositions of Korean properly. Therefore, as an attempt to solve this problem, models were

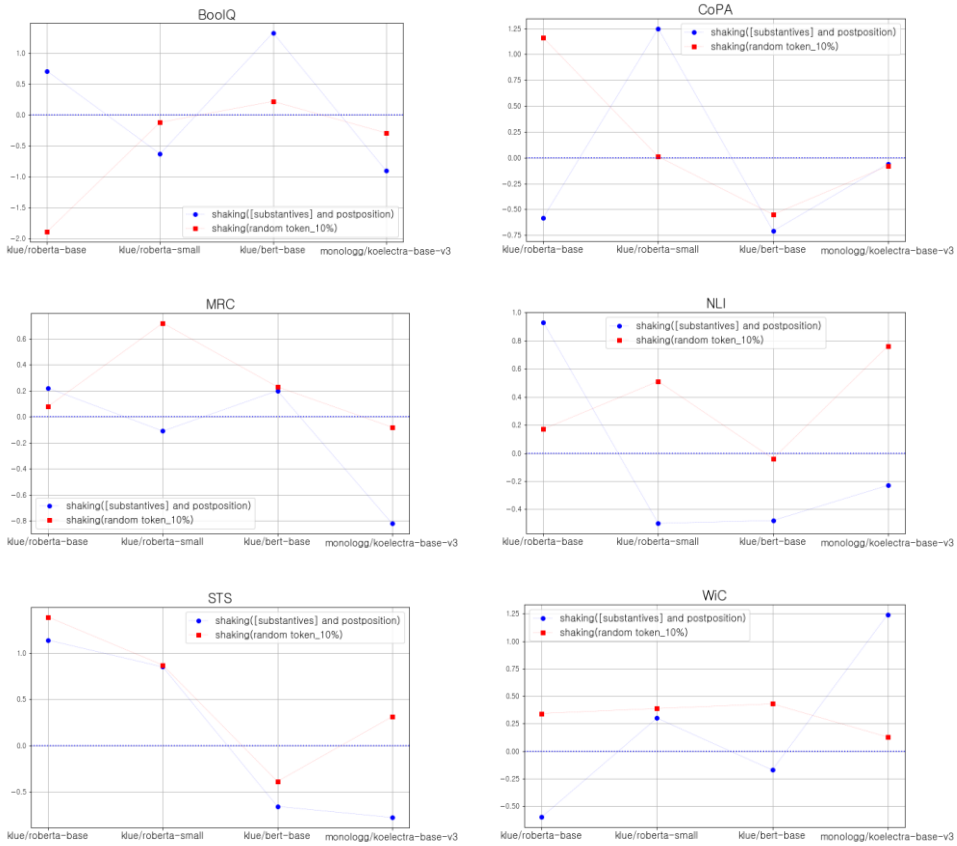
trained after making the positional encodings the same or combining the embeddings.

Since all the conditions except for the training method are the same, a relative comparison of the performances is more meaningful. However, in the case of the test set made by dividing the evaluation set in half, the result may be affected by the seed number for a random split. Since the performance depends on the seed number, making an accurate relative comparison takes work. Therefore the relative performance difference was compared only with the evaluation set without dividing it in half.

Chapter 5. Results and Analysis

Shaking at finetuning stage only

Figure 1 shows the results of seven downstream tasks for 4 HF pretrained models. The last image is an average of these results for each model. The marked positions on the graph represent the performance gain obtained by shaking attention scores compared to no manipulation case. Two methods of shaking the attention scores were applied. The one is to shake the substantives' attention, and the other is to do so for the random select tokens with a probability of 10%.



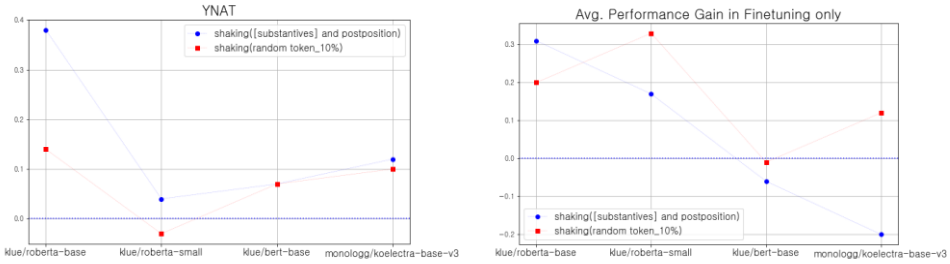


Figure 1: Test results of seven downstream tasks for 4 HF pretrained models

Looking at the results, we can find apparent differences in performance depending on the pretrained model, the downstream task, and the method of manipulating the attention scores. Better results from random select tokens are similar to the generalization effect caused by regularization. In any case, achieving high performance is only sometimes possible, and we must choose a proper model and method according to the downstream task. No pretrained model can be applied to various NLU tasks in common, and it is not good from a practical point of view. By the way, among these downstream tasks, no improvement is shown in YNAT, which classifies Yonhap News headlines into seven classes, even when various methods are combined. It is because the dataset has omitted words due to the nature of headlines, such as "포스코건설 11년 만에 더 샵 브랜드 로고 교체 (POSCO E&C's brand logo change in 11 years)". Therefore there are few postpositions, or the sentence is short, so the number of tokens for changing the attention scores is relatively small. Therefore, the performance gain or deviation obtained by manipulating the attention scores is tiny.

Shaking while pretraining

Six models were trained from scratch, and the evaluation losses of these

models are shown in Figure 2. The lowest loss can be obtained from the models of changing the substantives' attention scores with boost_factor 0.3 or -0.3. The case of normal training for baseline and changing the attention scores by randomly selecting tokens show a similar loss.

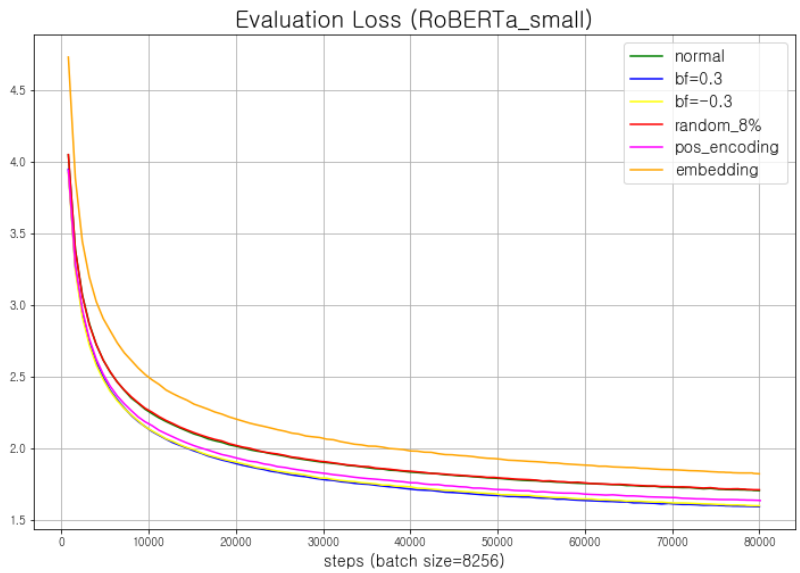


Figure 2: Evaluation losses of pretrained models

Figure 3 shows the results of six downstream tasks for the six pretrained models and an average of these results for each model. All marked positions are the relative performance difference with the baseline: a model was pretrained and finetuned without manipulating attention scores. The blue horizontal line corresponds to baseline performance. The red horizontal line represents the performance gain obtained by manipulating the attention scores of the substantives while finetuning with the normal model. As in the previous results, performance deviations depend on the pretrained model or downstream task. In most tasks except BoolQ, the performance of the pretrained model of boost_factor 0.3 can be improved considerably, even in case of

no manipulation in finetuning. In particular, the performance improvement in the CoPA task is remarkable.

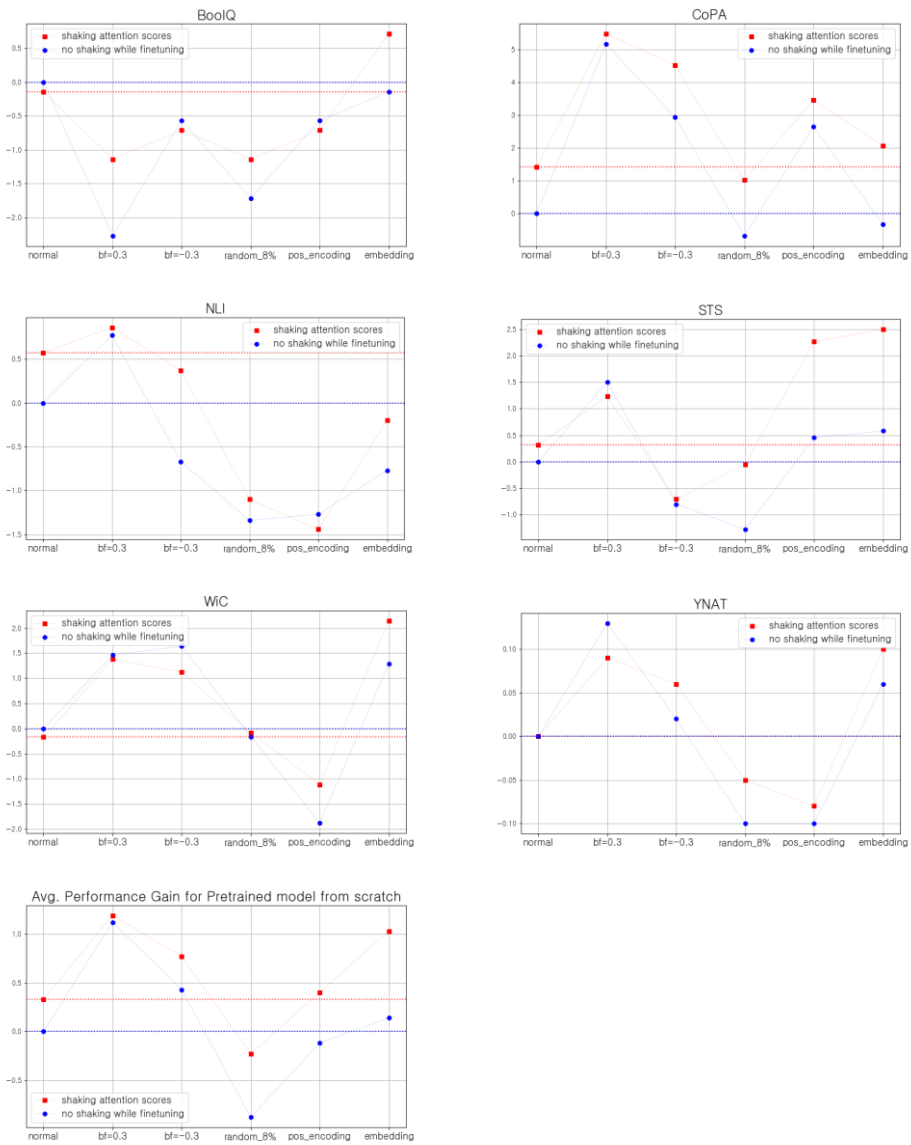


Figure 3: Test results of six downstream tasks for the models pretrained from scratch

Figure 4 shows the effect of the pretraining steps on performance. Each marked position represents the average additional performance compared to no manipulation case. For all tested downstream tasks, the

more pretrained steps, the higher the performance improvement.

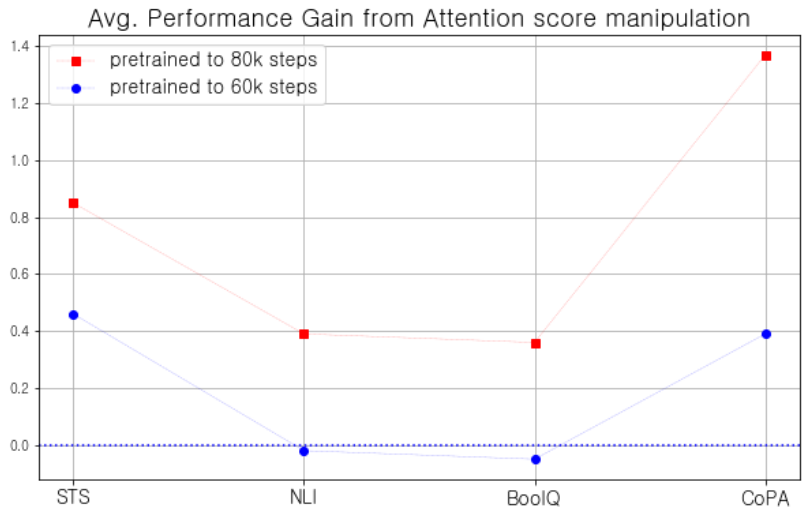


Figure 4: Effect of the maturity of the pretraining on performance

Chapter 6. Future work

This study is based on the premise that the current Transformer learning algorithm could not be able to learn Korean sufficiently, especially postpositions. Based on this, it was shown that additional NLU performance could be obtained by changing the attention scores of substantives or randomly selected tokens. However, it was confirmed that the performance variance mainly depended on the pretrained models or downstream tasks. The lack of analysis for the performance deviations is a challenge to be addressed in the future.

Through this process, the ultimate direction of this study is to develop a new Transformer learning algorithm that reflects the characteristics of Korean. In addition, of course, the PE algorithm also requires improvement because the word order can be varied more freely compared to English.

Chapter 7. Conclusion

Unlike English, postposition is very important for an agglutinative language like Korean. Word order in a sentence has a higher degree of freedom than in English, thanks to the significant role that postposition plays in the meaning of a sentence. Unfortunately, the current Transformer algorithm did not reflect the characteristics of Korean as an agglutinative language but perfunctorily applied the same as English. Therefore, the primary training method, which is predicting a target word using the given words, needs to improve to better suit the characteristics of Korean.

This study showed that NLU performance could be improved by manipulating the Transformer's attention scores by reflecting the characteristics of Korean. Research in this direction is hard to find in Korea, not to mention the English-speaking world. Therefore, this study is sufficient to say that it is the beginning of research to improve performance by directly manipulating attention, a crucial part of the Transformer. Performance deviations depend on the pretrained models or downstream tasks, and there is a limitation in analyzing them clearly. Nevertheless, it was shown that some models and tasks could benefit from performance gains by manipulating the attention of substantives. In addition, tests were conducted on cases in which positional encodings and token embeddings were manipulated from the pretraining stage. In all of these tests, the best performance can be obtained from the model trained from scratch by manipulating attention scores of substantives with a positive boost_factor.

Bibliography

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[2] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[4] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel SWeld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[6] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*,

pages 8821–8831. PMLR.

[7] Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: getting inside bert's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

[8] Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

[9] Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[11] Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. 2020. Improve transformer models with better relative position embeddings. *arXiv preprint arXiv:2009.13658*.

[12] Shaw, P., Uszkoreit, J., and Vaswani, A. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018.

[13] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information*

processing systems, pages 3104–3112, 2014.

[16] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational linguistics.

[17] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

[18] Hao, Y., Dong, L., Wei, F. and Xu, K. 2021. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*. 35, 14 (May 2021), 12963-12971.

[19] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*.

[20] Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183*.

[21] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[23] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D.

Manning. 2020a. Electra: Pretraining text encoders as discriminators rather than generators. In *ICLR*.

[24] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh et al., Klue: Korean language understanding evaluation, *arXiv preprint arXiv:2105.09680*, 2021.

Abstract in Korean

한국어는 영어와 분명히 다른 특성을 갖고 있지만 이를 Transformer에 반영하여 한국어에 보다 부합하는 새로운 모델을 찾는 시도는 그리 충분하지 않다. 본 연구에서는 한국어 특성 중에 특히 조사의 역할에 주목한다. 조사 덕분에 영어와 같은 굴절어에 비해 문장 내 단어 순서의 자유도가 높은 교착어라는 특성을 반영하여 Transformer의 attention score 계산 방법의 변경을 제안한다. 본 연구는 한국어와 같은 교착어에서 매우 중요한 역할을 하는 조사가 현재의 Transformer에서는 충분히 학습되기 어렵다는 가설에 바탕을 둔다. 한국어에서 조사는 해당 체언과 쌍으로 묶이므로 문장 내의 다른 token에 비해 해당 체언을 좀더 attention하는 것이 타당해 보이지만 현재의 Transformer 학습 방법으로는 한계가 많다는 의미이다. 이에 조사-체언 간의 attention score를 인위적으로 변화시킴으로써 NLU(Natural Language Understanding) 관련 자연어 처리 task의 성능을 높일 수 있음을 보인다. 아울러 한글 특성을 반영한 새로운 학습 방법에 관한 연구에 자극이 될 수 있기를 기대한다.