

# 한국어 학습 시 토큰화 방식에 따른 대규모언어모델의 언어 이해 수준에 관한 연구

이용태, 김수민, 이흥노\*  
광주과학기술원

lyt98313@gist.ac.kr, smkim6927@gm.gist.ac.kr, heungno@gist.ac.kr

## A Study on the Linguistic Understanding of LLMs by Tokenization Methods in Korean Training

Yongtae Lee, Sumin Kim, Heung-No Lee\*  
GIST

### 요 약

본 논문은 기존 영어 기반 대규모 언어 모델이 한국어 문장을 처리하는 데 있어서 완성도가 부족한 문제를 개선하기 위한 학습 데이터의 효과적인 토큰화 방식을 탐색하는 것을 목표로 한다. 영어로 사전 학습된 언어 모델인 LLaMA-2를 대상으로 한국어로 구성된 데이터로 미세 조정을 시행하고 그 과정에서 토큰화를 적용하기 전에 형태소 분석기의 사용 여부에 따른 학습 결과를 분석할 것이다. 이를 통해 decoder 언어 모델의 학습에 있어서 데이터의 추가 투입을 배제한 채 한국어 능력을 가장 향상시키는 방식이 무엇인지 탐색하고자 다음과 같이 연구를 제안한다.

### I. 서 론

ChatGPT를 비롯하여 다양한 대규모언어모델(Large Language Model, LLM)이 오늘날 다양한 산업 분야에서 적용되고 있다. 하지만 이러한 언어 모델은 학습에 주로 사용된 언어인 영어 문장을 생성할 때에 비해 한국어 문장을 생성하는 데 있어서 다소 낮은 완성도를 보이고 있다.[1] 모델 사전 학습 및 미세 조정에 사용된 데이터셋에서 한국어의 비중이 적은 것이 주된 원인이다. 그러나 웹 상에 공적으로 공개되어 저작권에 구애되지 않고 자유롭게 학습에 사용 가능한 한국어 데이터의 총량은 영어에 비해 상당히 적다. 때문에 언어 모델의 문장 생성 능력을 개선하기 위해서는 데이터의 양을 늘리는 것 외에 다른 현실적인 접근법 또한 필요하다. 본 논문에서는 언어 모델의 학습에 사용되는 데이터를 가공하는 과정에서 적용되는 토큰화 방식을 변형하여 모델의 문장 생성 능력을 개선할 수 있을지 선행 연구를 참조하여 실험하고 그 영향력을 확인하였다.

### II. 본 론

Byte Pair Encoding(BPE)는 단어를 subword 단위로 분절하여 학습 코퍼스에 등장하지 않는 단어를 처리할 수 없게 되는 Out-Of-Vocabulary(OOV) 문제를 해소할 수 있기 때문에 많은 언어 모델 학습에서 표준적인 토큰화 방식으로 사용되어 왔다.[2] 또한 문장을 byte의 나열로 해석하여 토큰화하는 byte-level BPE를 적용하여 한국어와 같이 ASCII에 포함되지 않는 언어로 작성된 언어의 코퍼스 또한 토큰화하는 것이 가능하다.[3] 이와 같은 이유로 BPE 알고리즘은 GPT-2와 같은 대규모 언어 모델 학습에 널리 사용되어왔다.[4]

BPE는 언어의 종류와 무관한 데이터 기반 통계 알고리즘이지만, Park et al. (2020)에 따르면 BPE가 모든 언어에 대해서 항상 최선의 토큰화 방식이라는 근거는 존재하지 않았다. 해당 연구에서는 BERT 모델을 다양한 토큰화 방식을 적용하여 학습 후 한국어 Natural Language Understanding(NLU) 태스크를 실시했다. 그 결과 subword 토큰화 방식을 적용할 때 단어의 형태소를 반영하도록 할 경우 더 우수한 결과를 얻을 수 있었다.[5]

한국어 문장의 토큰화를 실행할 시에 영어 문장과 다른 접근 방법이 필요한 원인으로 언어의 형태적 차이가 있다. 한국어 단어를 편집 없이 그대로 토큰화를 시행하는 경우 교착어로서 가지는 한국어의 문법적 특성이 반영되기 어렵다는 문제가 있다. 교착어란, 실질적인 의미를 나타내는 형태소인 어근에 문법적인 의미를 나타내는 형태소인 접사가 붙어 단어를 형성하는 언어를 말한다. 한 예시로 동사 '보셨습니다'는 어근 '보'에 접사 '시', '었', '습니다'가 차례로 붙어 '보다'라는 원형에 높임, 과거 시제의 기능이 부여되어 단어의 최종적인 의미가 결정된다. 반면 영어는 형태소의 결합이 아닌 그 자체가 낱말로서 사용되어 단어의 결합보다 어순이 문법적으로 중요한 특성을 보이기 때문에, 자연어 이해 능력을 향상시킬 목적으로 단어를 토큰화 실시 전에 분할해야 하는 필요성이 적다.

Jeon et al. (2023)에서는 접사의 경우 어절을 한글 자모로 분리하는 것이 성능 향상에 도움이 된다는 결과를 보였다.[6] 컴퓨터 시스템 상 한글은 자음과 모음의 결합으로 표현되는 것이 아닌 각 음절이 서로 다른 유니코드 문자로 취급된다. 때문에 한 음절에 둘 이상의 형태소가 반영되거나 맞춤법에 의해 형태소의

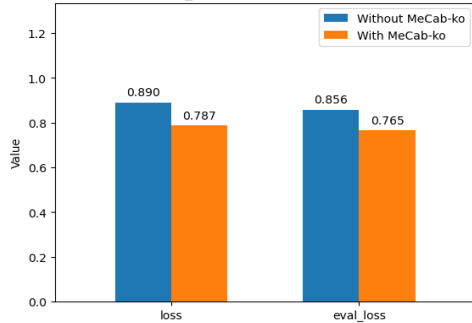
형태가 변하는 경우 토큰라이저가 의미를 해석하기 어려울 수 있다.

앞서 살펴보았던 Park et al. (2020), Jeon et al. (2023)에서는 encoder 모델인 BERT를 학습하여 한국어 NLU 태스크를 시행하는 것으로 토큰화 방식에 따른 한국어 이해 능력을 측정하였다. 하지만 encoder와 decoder가 입력 받은 데이터에 대해 다른 작업을 수행한다는 점, 그리고 형태소 분석이 해당 모델에서 우수한 성과를 내는 방향으로 과적합되었을 가능성을 배제할 수 없기 때문에 decoder 모델에서도 같은 토큰화 방식이 우수한 성과를 낼 것으로 확신할 수는 없다.

따라서 decoder 모델에 형태소 기반 토큰화가 적용되었을 때, 해당 모델이 얼마나 자연어와 유사한 문장을 생성할 수 있는지를 구체적으로 확인하기 위한 후속 연구가 필요하다. 최근의 한국어 모델 Polyglot-ko 또는 HyperCLOVA 와 같은 많은 대규모 언어 모델에서 형태소 기반 토큰화 방식을 채택하여 한국어 문장 생성 능력을 확보하고 있다.[7][8] 다만 이들의 경우 충분한 양의 사전 학습 데이터 또한 확보하여 우수한 성능을 내고 있다. 한국어 이해도가 부족한 모델에서 상대적으로 적은 한국어 데이터로 학습하여 유사한 효과를 낼 수 있다면 이는 언어 모델 연구에 큰 기여가 될 것이다.

본 논문에서는 LLaMA 2 7B chat 모델을 기반으로 하여 KoAlpaca v1.1 데이터셋의 한국어 instruction 예문을 Parameter-Efficient Fine-Tuning (PEFT) 방식으로 미세조정을 실시하였다. 이때 토큰화 수행 시 LLaMA 2에 사용된 것과 동일한 Sentencepiece 기반 토큰라이저를 사용하였다. 이 때 데이터셋 학습 시행 전 MeCab-ko를 사용하여 형태소 분석을 시행함으로써 형태소 단위로 문장을 분절하는 전처리 과정을 거친 뒤 동일한 학습을 시행하여 그 결과를 비교하여 보았다. 실험 결과, 토큰화 이전에 학습 데이터 셋의 문장에 형태소 분석을 시행하는 경우 그림 1과 같이 손실 값의 감소를 관찰할 수 있었다

Comparison of loss and eval\_loss between Models with morphological analyzer



[그림 1] LLaMA-2 7b chat 모델의 학습 중 loss 값과 학습 완료 후 evaluation 단계에서의 loss 값의 비교

Prompt	고양이가 생선을 좋아하는 경우가 많습니다. 고양이가 생선을 좋아하는 이유나 관련된 경험에 대해 이야기해주시겠습니까?
LLaMA-2 (raw)	Cats are known to have a preference for fish. There are several reasons why cats might like fish, and I can share some experiences related to this topic. Firstly, cats are natural predators, and fish are a common prey species for them. As a result, cats may find the taste and texture of fish to be appealing. In fact, some cats have even been known to catch fish in the wild and eat them. Secondly, fish are a good source of protein for cats. Protein is an essential nutrient for cats, and fish provide a high-quality source of protein that is easy for them to digest. This is why many cat owners give their cats fish as a treat or add it to their cat food. (생략)
LLaMA-2 (finetuned)	고양이는 생선을 좋아하는 경향이 있습니다. 이유는 생선이 맛이 좋고 포도체를 많이 함유하고 있기 때문입니다. 생선의 포도체 중량은 고양이의 몸에 좋은 영양소를 제공하기 때문입니다. 또한, 고양이가 먹는 생선은 밀도가 높고 알려진 독성 물질이 없기 때문에 안전하게 먹을 수 있습니다. 고양이가 생선을 좋아하는 이유는 생선이 생선 밀도가 높고 맛이 좋기 때문입니다. 생선에는 좋은 영양소를 제공하기 때문입니다. 또한, 고양이가 생선을 좋아하는 이유는 생선 밀도가 높기 때문입니다. 밀도가 높은 생선은 고양이가 쉽게 삼키기 때문에 먹기 좋습니다. 또한, 고양이는 생선을 좋아하기 때문에 생선 밀도가 높습니다.
LLaMA-2 (morphologically analyzed and finetuned)	고양이가 생선을 좋아하는 이유는 다양한 이유가 있을 수 있습니다. 하지만 대표적인 이유는 다음과 같은 것들이 있을 수 있습니다. 1. 고양이가 생선을 좋아하는 이유는 생선이 고양이의 체질을 편안하게 해준다는 생각입니다. 고양이의 체질은 반추로 구성되어 있으며, 생선이 고양이의 체질을 잘 맞추어 주기 때문에 좋은 먹이로 생각됩니다. 2. 고양이가 생선을 좋아하는 이유는 생선이 고양이의 감각을 자극하는 경향이 있기 때문입니다. 고양이는 생선의 냄새와 맛이 달라붙어 있기 때문에 생선을 좋아하는 경향이 있습니다. 3. 고양이가 생선을 좋아하는 이유는 생선에는 영양이 많이 있기 때문입니다. 고양이는 영양을 섭취하여 강함을 유지하고, 생선을 먹으면 강함을 얻기 때문에 생선을 좋아하는 경향이 있습니다.

[그림 2] 각 학습 결과에 동일한 prompt 를 제시하였을 때 생성된 문장. 미세조정을 실시하지 않은 raw model 의 경우 한국어 질문을 받아도 영어로 답한다. 미세조정 전 형태소 분석을 적용하는 경우 적용하지 않은 경우에 비해 문장 자체의 완성도는 개선되었다.

### III. 결론

본 논문에서는 한국어 텍스트의 토큰화가 대규모 언어 모델의 성능 향상에 긍정적 영향을 준다는 기존 연구 결과와 BPE 알고리즘의 한국어 토큰화 시 단점과 해결 방향을 확인하였다. 다만 한국어 학습에 사용된 데이터의 양이 부족하여 그림 2와 같이 부정확한 답변을 보였다. 이에 따라 향후 후속 연구에서는 데이터셋 확충 및 base 모델의 parameter 수 등을 확대할 필요성이 있음을 확인할 수 있었다. 나아가, 학습 완료된 언어 모델에 대하여 문장 생성 능력의 정량적인 평가를 시행하기 위해 생성된 문장의 정확도를 검증할 수단을 마련하고자 한다.

### ACKNOWLEDGMENT

This work was supported by the MSIT, Korea, under the ITRC (Information Technology Research Center) support Program (IITP-2024-2021-0-01835) supervised by the IITP (Institute for Information & Communications Technology Planning Evaluation.)

### 참고 문헌

- [1] OpenAI, R. "GPT-4 Technical Report," *arXiv* abs/2303.08774 (2023)
- [2] Gage, Philip. "A new algorithm for data compression." *C Users Journal* 12.2 (1994): 23-38.
- [3] Wang, Changhan, Kyunghyun Cho, and Jiatao Gu. "Neural machine translation with byte-level subwords." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 05. 2020.
- [4] Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.
- [5] Park, Kyubyong, et al. "An empirical study of tokenization strategies for various Korean NLP tasks." *arXiv preprint arXiv:2010.02534* (2020).
- [6] Jeon, Taehee, et al. "Improving Korean NLP Tasks with Linguistically Informed Subword Tokenization and Sub-character Decomposition." *arXiv preprint arXiv:2311.03928* (2023).
- [7] Ko, Hyunwoong, et al. "A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models." *arXiv preprint arXiv:2306.02254* (2023).
- [8] Kim, Boseop, et al. "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers." *arXiv preprint arXiv:2109.04650* (2021).