Journal of the KNST

2023; Vol.6, No.2; pp. 201-206

ISSN: 2635-4926 https://doi.org/10.31818/JKNST.2023.06.6.2.201

Check for updates

Received: 2023/05/25
Revised: 2023/06/12
Accepted: 2023/06/29
Published: 2023/06/30

*Corresponding Author:
Dooyoung Kim
Dept. of Artificial Intelligence, Republic of Korea Naval Academy
Jungwon-ro, Jinhae-gu, Changwon-si, Gyungsangnam-do, 51704, Republic of Korea
Tel: +82-2-907-5246
E-mail: dykim07@navy.ac.kr

MIL-BERT: 군사 도메인 특화 한국어 사전학습 언어모델

MIL-BERT: Military Domain Specialized Korean Pre-trained Language Model

허희순[1], 윤창민[1], 유영하[1], 용석현[1], 김두영[2*]

[1]해군사관학교 사이버과학과 사관생도
[2]해군소령/해군사관학교 인공지능학과 부교수

Hee-Soon Heo[1], Chang-Min Yoon[1], Young-Ha Ryu[1], Seok-hyun Yong[1], Dooyoung Kim[2*]

[1]Midshipman, Dept. Cyber Science, Republic of Korea Naval Academy
[2]LCDR, ROK Navy/Associate Professor, Dept. of Artificial Intelligence, Republic of Korea Naval Academy

Abstract

본 논문에서는 추가 사전학습을 통한 군사 도메인에 특화된 BERT 모델을 제안한다. 기존 BERT 모델은 범용 코퍼스로 학습되어 특징 도메인에서의 활용에 최적화되어 있지 않다. 모델 학습을 위해 국방일보와 군사뉴스로부터 110만 개의 군사 문장과 6,900개의 군사용어를 수집하여 코퍼스를 구축하였다. 이후, 토크나이저를 구축하고 MLM학습을 통해 모델을 훈련했다. 또한, 성능 평가를 위해 MIL-BERT와 기존 한국어 BERT 모델인 KcBERT와 KoBERT 간의 군사 문장 분류 실험을 진행했다. 실험 결과, MIL-BERT가 정확도 측면에서 2 % 우수한 성능을 보였다.

In this paper, we propose a specialized BERT model that is tailored to the military domain through additional pre-training. Existing BERT models are trained on generic corpora and are not optimized for specific domains. To address this limitation, we collected 1.1 million military sentences and 6,900 military terms from military news to construct a corpus for model training. Subsequently, we developed a tokenizer and trained the model using masked language modeling (MLM). To evaluate the performance, we conducted military sentence classification experiments comparing MIL-BERT with existing Korean BERT models, KcBERT and KoBERT. The experimental results showed that MIL-BERT outperformed the other models with a 2 % higher accuracy.

Keywords

자연어처리(Natural Language Processing),
군사 언어(Military Corpus),
사전학습 언어모델(Pre-trained language model),
문장 분류(Sentence Classification),
심층 학습(Deep Learning)

1. 서론

자연어처리(natural language processing, NLP)는 인간의 언어를 해석, 조작 및 이해하는 능력을 모사하기 위한 기계학습 기법이다[1]. 다양한 NLP 기법 중 사전학습 언어모델(pre-training language model)[2]은 가장 일반적인 기법으로, 대량의 말뭉치를 비지도 학습함으로써 자연어 지식을 획득한다.

대표적인 사전학습모델 중 하나인 BERT(bidirectional encoder representations from transformers)[3]는 트랜스포머(transformer) 구조를 이용하여 문장 내 단어 간의 상호관계를 모사한 최초의 모델로, 이후 번역을 비롯하여 문서 요약, 챗봇과 같은 질의응답, 자연스러운 문장의 생성 및 문서의 분류 등 다양한 자연어처리 작업에 활용되고 있다. 실제로 BERT를 기반으로 미세 조정을 수행한 모델이 자연어처리에서 인간보다도 더 높은 정확도를 보여주었다[3].

이러한 장점에도 불구하고, 대부분의 BERT 모델은 범용 말뭉치로 학습되어 OOD(out-of-distribution)에 해당하는 의료, 법률, 금융 등의 특정 분야, 즉 특수 도메인에서는 취약한 면모를 보인다[4,5]. 이는 해당 모델들이 위키 문서, 웹 페이지, 일반 뉴스 등으로 대부분 구성된 범용적인 말뭉치를 기반으로 학습되어 특정 도메인에서의 활용에 최적화되어 있지 않기 때문이다. 이러한 도메인 특화 능력의 부재

© 2023 Korea Society for Naval Science & Technology

로 발생할 수 있는 가장 일반적인 현상으로는 (1) OOV (out-of-vocabulary) 발생으로 인한 언어 이해 부족 (2) 특정 도메인에서만 사용되는 단어 및 유의어 관계에 대한 이해 부족 등이 있다[3]. 이러한 문제를 해결하기 위해서는 특정 전문 분야의 문맥 정보 학습을 통해 BERT 언어모델의 가중치를 갱신하는 작업이 필요하다. 따라서 일반적인 단어를 학습한 범용 BERT 모델에 특정 도메인 지식만 추가하는 추가 사전학습(further pre-training)방식을 통해 성능을 향상하는 다양한 연구들이 진행되었다[6-8].

자연어처리 기술의 적용은 민간영역뿐만 아니라 군사영역으로도 점차 확대되고 있다. 2022년 3월에는 러시아 병사들이 전쟁 중 무선으로 나눈 대화 내용이 SNS를 통해 공개됐다. 자연어처리 알고리즘이 발화를 자동으로 기록, 번역, 분석하는 데 활용된 것이다[5]. 또한, 국방기술진흥연구소의 '미래국방 2030 기술전략'에서는 자연어처리 기술을 미래 핵심 AI 기술로 언급하며, 이를 통해 지휘 결심의 단계 중 지시기반 임무 계획의 발전이 가능함을 제시하였다[8]. 따라서 군사 언어의 해석이 가능한 자연어처리 모델의 연구 및 개발이 필요하다.

본 연구에서는 추가 사전학습을 통한 '군사 도메인 특화 BERT 모델을 제안한다. 기존 범용 한국어 BERT 모델인 KcBERT 토크나이저(tokenizer)에 국방일보와 군사용어 사전에서 수집한 대량의 군사 도메인 말뭉치를 이용하여 군사 도메인 지식을 추가로 학습시킨다. 이후, 사전 학습된 군사 도메인 특화 BERT 모델의 우수성을 입증하기 위해 기존의 범용 한국어 BERT 모델과 군사 문장 분류 비교를 시행하였다.

이 연구의 기여는 다음과 같다. 첫 번째, 군사 도메인에 특화된 사전언어학습모델을 제안하였다. 두 번째, 사전언어학습을 위한 군사용어 학습 데이터를 구축하였다. 세 번째, 구축된 학습 데이터를 이용하여 최초의 한국어 군사 도메인 특화 BERT 모델을 구현하였다. 네 번째, 구현한 모델의 다양한 국방 분야 활용 방안을 제시하였다.

논문의 구성은 다음과 같다. 2장은 연구에 사용된 사전학습 언어모델인 BERT와 도메인 특화 학습 기법에 관해 설명한다. 3장은 본 연구가 제안한 모델의 구조와 학습 방법에 설명하고, 4장은 모델의 성능 평가 실험에 관한 결과를 서술하였다. 마지막 5장에서는 결론 및 향후 연구를 위한 활용 방안에 관해 서술하였다.

2. 관련 연구

2.1 BERT

BERT는 구글에서 공개한 사전 훈련된 모델로, 양방향 문맥을 이해하고, 사전훈련과 미세 조정(fine-tuning) 과정을 통해 다양한 NLP 작업에서 높은 성능을 발휘한다[3].

BERT는 MLM(masked language modeling) 방식을 통해 주변 단어들의 양방향 관계를 학습한다. 이는 주어진 문장 내의 단어를 [MASK] 토큰으로 변환하여 주변 맥락을 통해 토큰이 된 단어를 예측함으로써 양방향 정보를 모두 반영한 텍스트를 학습할 수 있게 한다.

BERT의 높은 성능은 특유의 임베딩(embedding) 방식에 기인한다. 임베딩은 토큰(token) 임베딩, 세그먼트(segmentation) 임베딩, 위치(position) 임베딩으로 나뉜다. 토큰 임베딩은 입력 텍스트의 단어나 형태소를 고정된 차원의 벡터로 변환하는 과정이다. 이를 통해 모델이 단어를 이해하는 데 도움을 준다. 세그먼트 임베딩은 입력 텍스트에서 각 토큰이 속한 문장을 구분하는 역할을 한다. 위치 임베딩은 각 토큰의 위치정보를 포함한다. 이는 트랜스포머 구조의 경우 입력 토큰이 독립적으로 처리되어 단어 순서와 문맥 정보를 이해하기 위해서는 위치정보가 필요하기 때문이다.

이렇게 학습된 BERT는 분류와 객체 인식 등의 다운스트림(down stream) 과업에 활용할 수 있다[3]. 다운스트림 과업을 진행하기 위해서는 미세 조정을 통해 사전 훈련된 모델을 작업 관련 데이터를 사용해 추가로 학습시키는 과정을 거쳐야 해당 모델을 더 효과적으로 활용할 수 있게 된다. 미세 조정은 하위 분석 과제 수행을 위해 사전학습 언어모델의 매개변수(parameter)를 재학습을 통해 미세하게 조정하는 방식이다[3].

BERT는 영어를 기초로 한 다국어(multilingual) 모델로 제작되었기에, 한국어 처리 성능에는 한계가 존재한다[9]. 이러한 한계를 극복하기 위해 SKT Brain의 KoBERT[12]는 기존의 BERT 모델에 한국어 뉴스 2천만 문장을 추가 학습시켜 한국어 문장의 분석 성능을 향상시켰다.

2.2 도메인 특화 학습

사전학습 모델을 사용 목적에 맞게 조정하지 않으면

성능이 저하된다[4]. [4]에서는 다양한 도메인 데이터에서 전이학습(transfer learning)을 수행하여 모델을 비교하여 성능이 제한된다는 연구 결과를 도출하였고, [5]에서는 도메인 특화 사전학습이 모델의 성능에 미치는 영향을 비교 분석하였다. 또한 [5]에서는 도메인 특수성이 낮으면 보편적인 언어모델을 사용하는 것이 효과적일 수 있으나, 도메인 특수성이 높으면 해당 도메인에 특화된 언어모델을 사용하는 것이 효과적이라고 말하며 도메인 특화 언어모델의 필요성을 언급하였다.

앞서 언급된 도메인 특화 언어모델 구현을 위해 다양한 연구가 진행되고 있다. [13]에서는 자연어처리를 위한 연속학습(countinaul learning) 개념을 제시하였다. 이번 연구에서 사용하는 추가 사전학습은 연속학습을 구현하기 위한 기술 중 기존의 데이터를 남겨두고 새로운 데이터를 학습시키는 방식과 유사하다. [14]에서는 추가 사전학습의 유용성과 도메인 특화 데이터 세트의 중요성을 제시하고, 한국어 처리 분야에서 추가 사전학습 모델의 적용 가능성을 연구했다.

한편, NLP 분야에서 새로운 데이터에 대한 일반화 성능이 떨어지는 과적합(over-fitting)을 방지하는 방법에 관한 연구[15], 모델의 학습 효율을 높이기 위한 데이터 전처리[16], 딥러닝 개념 정리[17]에 관한 연구도 진행되었다. [16]은 딥러닝 모델 자체에 관한 연구가 아닌, 어떻게 데이터를 전처리해야 모델의 학습 성능을 높일 수 있는지 그 방향과 방법을 제시했다. 또한, BERT에서 발생하는 OOV 발생을 줄이는 방법도 제시하였다. [16]에서는 딥러닝 개념 중 혼용되고 있는 용어들에 대한 개념을 정리하고, BERT에서 추가 사전학습이 가능한 이유를 설명하고 있다. [17]은 과적합 현상을 방지하기 위한 조기 멈춤(early stopping) 기법에 관해 설명하고, 적절한 에포크(epoch) 수를 찾는 방법을 연구하였다.

3. 제안 모델

3.1 전체 학습 구조

Fig. 1은 본 연구에서 제안하는 군사 도메인에 특화된 BERT 모델의 전체 학습 구조이다. 먼저, 국방 도메인 학습 데이터 구축을 위해 국방일보와 한국군사문제연구원 등 군사뉴스 자료에서 대량의 기사를 크롤링하였다. 또한 군사 용어사전에서 6,900개의 군사용어를 추출하였다. 이 데이터를 기반으로 군사 문장으로 구성된 코퍼스 데이터를 구축하였고, word piece 토큰화를 통해 18,000개의 군사 도메인 vocab을 추출하였다. 이후 이 vocab을 KcBERT의 기존 토크나이저에 추가학습시켜 군사용어 토크나이저를 구축하였다. 이후 MLM 학습으로 모델의 사전학습을 진행하여, 군사 도메인에 특화된 MIL-BERT 모델을 구축하였다.

3.2 군사용어 토크나이저 구축

군사 도메인에 특화된 BERT를 사전학습(pre-training)시키기 위해서는 군사용어가 포함된 토크나이저 구축이 필요하다. BERT는 사전에 구축된 토크나이저를 통해 단어를 분절하므로, 특정 도메인에 특화된 BERT 말뭉치에 존재하지 않는다면 이를 여러 개의 하위 단어로 분절하여 전문어가 가진 고유한 의미가 손실된다[1].

군사 문장 데이터는 '국방일보'를 포함한 다양한 신문사의 국방 분야 기사와 군사용어사전의 단어들을 크롤링을 통해 수집한다. 이때, 크롤링한 문장 데이터에는 비문이나 원하지 않는 정보가 포함된 문장이 존재하므로 적절한 전처리가 필요하다. 데이터 정제를 마친 군사용어 코퍼스로 word piece 토큰화를 수행하여 군사 도

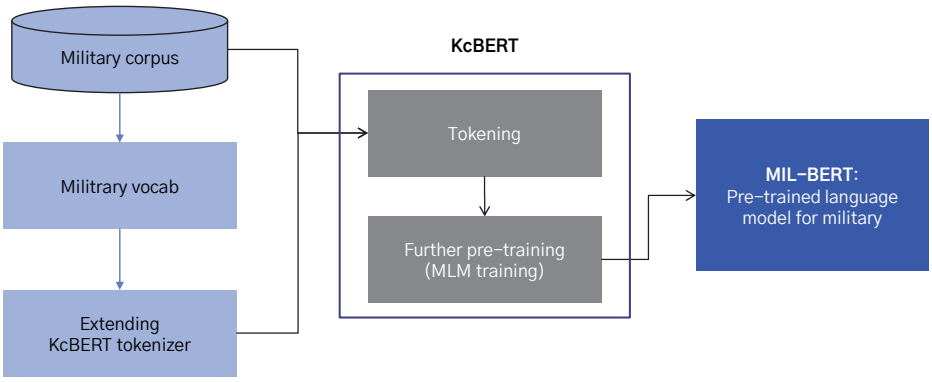

Fig. 1. Overall architecture of the proposed model

메인의 vocab을 추출한다. 이후 KcBERT 토크나이저에 추가 학습시켜 군사용어 토크나이저를 구축한다.

3.2 모델구조 및 학습

MIL-BERT는 트랜스포머 블록 12개, 어텐션 헤드(attention head) 12개, 그리고 은닉층 768개로 구성되며, 처리 가능한 문장의 최대 길이는 128단어로 한정하였다.

본 절에서는 이전 단계에서 구축한 군사영어 토크나이저를 활용하여 2단계로 나누어 MLM 학습을 진행한다. 먼저, 군사용어 토크나이저를 사용하여 군사 corpus를 토큰화하였다. 일반 BERT 토크나이저로는 "탄도미사일"과 "상륙작전"을 인식하지 못하여 더 작은 단위로 분절하지만, MIL-BERT의 군사용어 토크나이저는 이러한 용어를 올바르게 인식하여 토큰화한다.

다음으로, 분절된 군사용어 문장을 이용하여 BERT의 학습 방식 중 하나인 MLM을 활용한 추가 사전학습을 수행하였다. MLM은 입력 문장에서 임의의 단어를 마스킹하고, 모델이 이를 예측하도록 하는 방식이다. 예를 들어, "딥러닝"과 "합동 상륙작전"이 포함된 문장에서 "합동 상륙작전"을 마스킹하고 모델이 이를 정확하게 예측하도록 학습한다. MIL-BERT는 이러한 학습을 통해 군사용어의 의미정보와 지식을 충분히 정확하게 표현할 수 있도록 학습된다. 이를 통해 군사 분야의 텍스트 데이터를 처리하고 이해하는 능력이 강화된다.

4. 실험 및 결과

4.1 군사용어 토크나이저 구축 결과

코퍼스 데이터 수집을 위하여 국방일보의 국방 분야 90,000개 기사와 한국군사문제연구원의 군사뉴스 스크랩 자료의 12,869개의 기사를 크롤링하였다. 크롤링 결과 110만 개의 문장을 수집했다. 이후 110만 개의 텍스트 데이터에 대하여 데이터 전처리를 시행했다. 비문 제거를 위해 단어의 개수가 5개 이하인 줄은 모두 삭제하였다. 이후 문장 중간에서 줄바꿈이 발생하지 않도록 줄바꿈을 제거하였다. 다시 문장 단위로 나누기 위해 마침표, 느낌표, 물음표를 기준으로 문장을 나누었다.

비문처리와 문장 단위 분할이 끝난 이후 군대와 관련이 없는 문장을 삭제하기 위해 군대와 관련이 적은 단어 리스트를 작성하여 해당 단어를 포함하고 있는 문장들은 삭제하였다. 또한, 군사 용어사전에서 크롤링한 군사용어 6,900개와 일반화 성능을 높이기 위한 범용 도메인 용어 8,000개를 추가하여 최종 코퍼스 데이터를 구축했다. 그 결과로 군사 분야에 특화된 18,000개의 vocab을 추출하였다. 이 vocab을 KcBERT 토크나이저 30,000개에 추가로 학습시켜, 최종적으로 42,492개의 MIL-BERT 토크나이저를 구축했다.

Fig. 2는 MIL-BERT 토크나이저의 일부를 표현한 그림이다. '탄도미사일', '상륙작전' 등의 군사용어가 잘 포함되어있음을 확인 가능하다.

Fig. 2. Example of tokenizer

Fig. 3는 KcBERT의 토크나이저와 본 연구에서 구축한 군사용어 토크나이저로 실제 군사 문장에 대하여 토큰화를 실시한 결과이다. 이를 통해 군사용어 토크나이저는 '침투', '제대', '지휘소' 등과 같은 군사용어가 분절화되지 않고 잘 보존됨을 확인하였다.

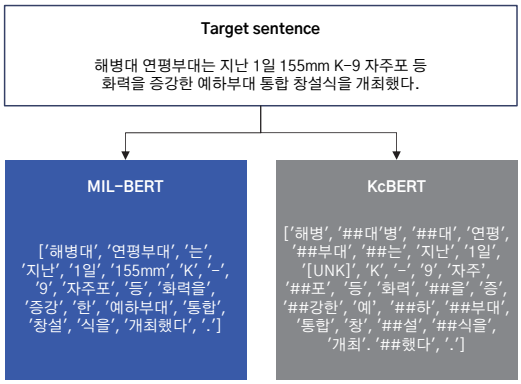

Fig. 3. Comparison of tokenization results

4.2 모델 학습

국방일보와 한국군사문제연구원에서 크롤링한 문장들을 6:2:2의 비율로 나눠서 각각 훈련용, 검증용, 그리고 평가용 데이터로 이용하였다. 따라서 훈련에 사용된 문장 개수는 455,839개이고, 검증 및 평가에 사용된 문장의 개수는 151,948개이다.

모델 학습에 이용된 문장은 총 759,735개이다. 학습에는 스케줄러를 사용했으며, 초기 학습률은 5e-5로 설정했다. 학습은 총 3 epoch 수행했으며 mini batch 크기는 32로 설정했다.

모델 학습 결과, training loss와 validation loss는 Fig. 4와 같이 나타났다. 이후, early stopping을 이용하여 학습 단계에서 손실률이 가장 낮은 에포크를 선택하여 최종모델을 만들었다.

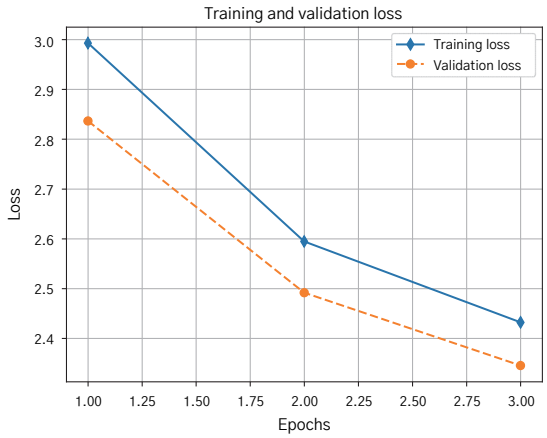

Fig. 4. Training and validation loss

4.3 군사 문장 분류 적용 및 성능 평가

본 절에서는 본 연구에서 제안한 Mil-BERT의 성능을 평가하기 위하여 한국어 BERT 모델인 KcBERT와 KoBert 간의 성능 비교 실험을 진행한다. 성능 비교는 군사 문장의 2개의 하위분류를 구분하는 binary text classification 실험을 진행한다. 이를 위해 '국방과학기술 용어사전' 중 '함정', '항공' 카테고리의 예문들을 추출하여 분류 실험을 수행하였다. Table 1은 군사 문장을 하위 분류한 예시이다.

성능 평가에는 학습 데이터와 별도의 2,149개의 라벨링된 문장을 6:2:2의 비율로 나눠서 각각 훈련용, 검증용, 그리고 평가용 데이터로 이용하였다. 이때, '함정'에 속하는 문장은 0, '항공'에 속하는 문장은 1로 라벨링하여 실험을 진행하였다. 5 epoch 학습을 진행하였고, 평가지표로는 precision, recall, f1-score, accuracy를 측정하였다.

Table 1. Examples of military and non-military sentence classification

Class	Sentences
Warship	잠수함의 존재를 나타내는 몇 가지의 정보가 있으나 높은 식별 등급을 부여하기에는 불충분한 수중 접촉물.
Aircraft	적의 항공기 및 유도탄을 공중에서 제거, 방어하기 위하여 구축한 대공 방어 체계.

Fig. 5와 Fig. 6는 제안 모델인 MIL-Bert와 기존의 KcBERT 모델에 대한 분류 학습의 epoch에 따른 training loss와 validation loss를 나타낸다. 테스트셋 평가는 가장 loss 값이 낮은 epoch을 채택하여 성능 평가를 진행하였다. 성능 평가결과는 Table 2와 같다.

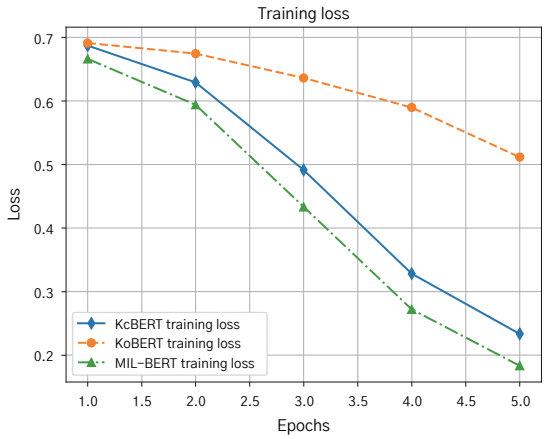

Fig. 5. Comparison of training loss

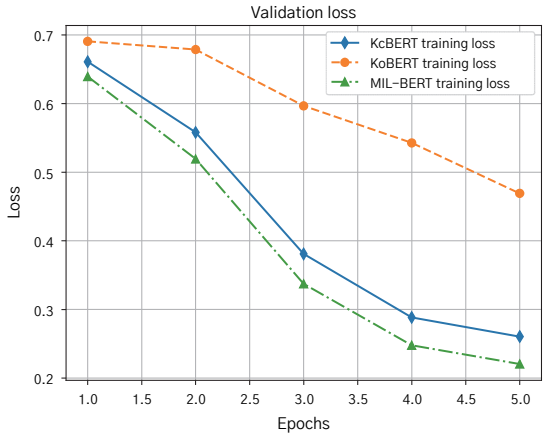

Fig. 6. Comparison of validation loss

Table 2. Performance evaluation results

Model	Class	Precision	Recall	f1-score	Accuracy
MIL-BERT	Warship	**0.96**	**0.91**	**0.94**	**0.94**
	Aircraft	**0.92**	**0.97**	**0.94**	
KcBERT	Warship	0.94	0.90	0.92	0.92
	Aircraft	0.91	0.95	0.93	
KoBERT	Warship	0.95	0.88	0.92	0.92
	Aircraft	0.90	0.96	0.93	

군사 문장 분류 실험 결과, MIL-BERT가 KcBRT와 KoBERT보다 모든 성능 지표면에서 우수한 성능을 보였다. 특히 accuracy에서는 타 모델보다 2 % 정도 뛰어났다. 따라서, 본 연구에서 제안하는 군사 도메인 특화 BERT 모델인, MIL-BERT가 군사 문장 분석에 효과적으로 적용할 수 있음을 확인하였다.

5. 결론

본 논문에서는 군사 분야 특화 학습을 수행하고 이를 기존의 BERT 모델과의 비교 실험하여 새로운 모델의 성능을 검증하였다. 우선 국방일보와 군사뉴스에서 110만 개의 군사 문장과 6,900개의 군사 단어로 구성된 코퍼스 데이터를 확보했다. 이후 word piece 토큰화 단계를 거쳐 18,000개의 군사용어 vocab으로 추출하였고, KcBERT 토크나이저에 추가 학습시켰다. 최종적으로는 42,492개의 MIL-BERT 토크나이저를 구축했다. MLM 학습을 통하여 추가 사전학습을 했고, 학습 단계에서 손실률이 가장 낮은 에포크를 선택하여 최종모델을 구축하였다. 이후, 성능 평가를 위해 군사 문장 분류 실험을 진행한 결과 MIL-Bert가 KcBERT와 KoBERT보다 모든 성능 지표에서 우수한 성능을 보였다. 특히 accuracy는 2 % 뛰어난 성능을 보였다.

본 논문에서는 군 내에서 공연히 사용되는 약어나 은어에 대한 학습이 부족하므로, 향후 이에 관한 추가적인 연구가 필요하다. 특히 군사용어의 경우 사회에서 흔히 사용하는 용어와 동음이의어 관계인 경우가 있어 이를 분류하는 연구를 진행하고, 학습용 데이터의 양을 늘려 정확도를 개선해 나갈 예정이다.

참고문헌

[1] Dpectrum. “What is Natural Language Processing (NLP) in the field of Artificial Intelligence?” https://dpectrum.app/blog/89
[2] 김동규, 박장원, 이동욱, 오성우, 권성준, 이인용, 최동원. KB-BERT: 금융 특화 한국어 사전학습 언어모델과 그 응용. 지능정보연구, Vol. 28, No.2, 2022, pp.191-206.
[3] Devlin, Jacob, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv preprint arXiv:1810.04805 (2018).
[4] Rogers, Anna, et al. ”Investigating Transferability in Pretrained Language Models.” arXiv preprint arXiv:2009.13393 (2020).
[5] 한민아, 김윤하, 김남규. (2022). 도메인 특수성이 도메인 특화 사전학습 언어모델의 성능에 미치는 영향. 지능정보연구, Vol. 28, No. 4, 251-273.
[6] Beltagy, Iz, Kyle Lo, and Arman Cohan. ”SciBERT: A Pretrained Language Model for Scientific Text.” arXiv preprint arXiv:1903.10676 (2019).
[7] Lee, Jinhyuk, et al. ”BioBERT: A Pretrained Biomedical Language Representation Model for Biomedical Text Mining.” Bioinformatics Vol. 36, No.4 (2020): pp.1234-1240.
[8] Araci, Dogu. ”FinBERT: Financial Sentiment Analysis with Pretrained Language Models.” arXiv preprint arXiv:1908.10063 (2019).
[9] Chalkidis, Ilias, et al. ”LEGAL-BERT: The Muppets Straight Out of Law School.” arXiv preprint arXiv:2010.02559 (2020).
[10] AI Times. ”AI, 군사작전에 교묘히 스며들기 시작.” Accessed July 13, 2022. https://www.aitimes.com/news/articleView.html?idxno=145765.
[11] 국방기술진흥연구소. “미래국방 2030 기술전략.” 2022
[12] SKTBrain. “Korean BERT Pretrained Cased (KoBERT).” https://github.com/SKTBrain/KoBERT.
[13] Jihye Lee, Hyeonmin Ha, Byung-Gon Chun. (2022). Survey on Recent Continual Learning Studies in NLP. 한국정보과학회 학술발표논문집, pp. 999-1001.
[14] Jaemin Lee, Younggyun Hahm, Donggyu Lee, and Hwanjo Yu. A Further Pretrained Language Model with Domain-specific Corpora for Question Answering in Korean. (2021). arXiv. 2104.06323
[15] Collobert, Ronan, et al. “Natural Language Processing (Almost) from Scratch.” Journal of Machine Learning Research Vol.15 (2014): pp. 335-366.
[16] 서혜진, 신정아.(2020).딥러닝을 활용한 감정 분석 과정에서 필요한 데이터 전처리 및 형태 변형. 영어학, Vol. 20, pp. 42-63.
[17] Yangoos57. ”Fine-Tuning with Pretrained Models.” https://yangoos57.github.io/blog/DeepLearning/paper/Finetuning/.