Journal of KIIT. Vol. 22, No. 1, pp. 51-59, Jan. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.1.51

Mil-BART : 국방 표준 자연어 데이터셋을 활용한 군사 도메인에 특화된 사전학습 언어모델

고관우*[1], 김한석*[2], 이수진**

Mil-BART : Military Domain Specialized Pre-trained Language Model using Defense Standard Natural Language Dataset

Gwan-Woo Goh*[1], Han-Seok Kim*[2], and Soo-Jin Lee**

요 약

특수 전문 분야에서 AI 기술 적용에 대한 연구는 전세계적으로 매우 활발하게 진행되고 있다. 그러나 국방 분야는 전문성과 보안이라는 높은 장벽 때문에 아직 많은 연구가 진행되지 못하고 있다. 본 연구에서는 국방 분야 AI 개발 여건 조성을 위해 구축을 추진 중인 국방 표준 자연어 데이터셋을 처음으로 적용해 군사 도메인에 특화된 Mil-BART 모델을 제안한다. Mil-BART 모델은 범용 말뭉치를 사전학습한 언어모델 BART의 토크나이저에 군사교범에서 추출한 토큰 49,107개와 국방논단에서 추출한 토큰 55,350개를 추가하여 구축하였다. Mil-BART의 성능 평가를 위해 군사 문장(군사교범 및 국방논단)과 비군사 문장(일반 뉴스)에 대한 이진분류 및 다중분류 실험을 진행한 결과, Mil-BART의 이진분류 재현률과 F1-score가 BART보다 뛰어난 것으로 나타났다. 특히 국방논단과 일반 뉴스에 대한 이진분류 재현률과 F1-score가 2%P 정도 향상됨을 확인하였다.

Abstract

Research on the application of AI technology in specialized domains is very active around the world. But, the defense domain has not been studied much due to the high barriers of expertise and security. In this study, we propose a Mil-BART model specialized in military domains by applying the defense standard natural language dataset, which is being built to create the environment for AI development in the defense domain. The Mil-BART model was constructed by adding 49,107 tokens extracted from the Military Manual and 55,350 tokens extracted from the Defense Argument to the tokenizer of BART, which is a pre-trained language model based on generic corpora. To evaluate the performance of Mil-BART, we conducted binary and multi-class classification experiments on military sentences(military manual and defense argument) and non-military sentences(general news). The experimental results show that Mil-BART outperforms BART in Recall and F1-score of binary classification, especially improving Recall and F1-score of the binary classification for defense argument sentences and general news sentences by about 2%P.

Keywords
defense standard natural language dataset, military domain, pre-trained language model, BART, Mil-BART, natural language processing, sentence classification

* 국방대학교 국방과학학과 국방과학석사
- ORCID[1]: https://orcid.org/0009-0005-5486-2794
- ORCID[2]: https://orcid.org/0009-0008-2365-4230
** 국방대학교 국방과학학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4117-407X

· Received: Oct. 26, 2023, Revised: Nov. 09, 2023, Accepted: Nov. 12, 2023
· Corresponding Author: Soo-Jin Lee
Department of Defense Science, Korea National Defense University, 1040 Hwangsanbeol-ro, Yangchon-myeon, Nonsan-si, Chungcheongnam-do, Korea
Tel.: +82-41-831-5378 Email: cyberkma@gmail.com

Ⅰ. 서 론

자연어 처리(NLP, Natural Language Processing)는 인간이 사용하는 언어를 컴퓨터가 이해하고 처리할 수 있도록 연구하는 분야이다. 이를 통해 컴퓨터는 텍스트나 음성을 이해하고, 그에 대한 응답을 생성할 수 있다. 언어모델(Language model)은 자연어 처리 분야에서 사용되는 핵심 개념으로 특정한 자연어의 시퀀스(Sequence)에서 다음에 나올 단어를 예측하는 모델을 의미하며, 미리 대규모의 데이터로 훈련된 언어모델을 사전학습 언어모델(Pretrained language model)이라고 한다.

자연어 처리 기술은 제반 분야에서 빠른 속도로 발전하고 있지만, 국방 분야는 전문성과 보안이라는 높은 벽 때문에 연구가 아직 미진한 상태이다. 또한, 민간에서 상용되는 언어모델을 군에 도입하기 위해서는 군사적 목적에 특화되고 국방 분야 AI 개발에 공통적으로 활용 가능한 표준화된 자연어 데이터셋이 필요하다. 이러한 문제를 해결하기 위해 현재 육군본부 정보화기획참모부와 민간전문업체가 협력하여 국방 표준 자연어 데이터셋 구축 사업을 진행하면서, 데이터셋의 유효성과 활용 가능성을 모색하기 위한 다양한 논의도 함께 진행되고 있다.

이에 본 연구에서는 국방 표준 자연어 데이터셋 중 전문가 검증이 완료된 텍스트 데이터셋 일부를 육군본부 승인하에 최초로 연구 목적으로 사용하여 이를 기존 BART(Bidirectional and Auto-Regressive Transformers)[1] 토크나이저(Tokenizer)에 추가시킨 Mil-BART 모델을 구축한 후, 군사-비군사 문장에 대한 분류 성능을 분석하였다. 즉 본 연구의 목적은 군사 용어 토큰을 추가로 학습한 언어모델이 범용 언어모델인 BART보다 군사 관련 용어를 더 잘 이해하여 군사 문장 분석에도 효과적으로 적용할 수 있는지의 여부를 확인함으로써 현재 구축 중인 국방 표준 자연어 데이터셋이 군 내 AI 학습용 데이터로 활용 가능한지를 검증하는 것이다.

이진분류 실험은 군사교범 문장 및 일반 뉴스 문장의 분류와 국방논단 문장 및 일반 뉴스 문장의 분류로 구분하여 실험을 진행하였다. 다중분류 실험은 각 군별 군사교범과 일반 뉴스 문장을 대상으로 진행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 BART 모델과 도메인 특화 학습에 관한 선행연구를 정리한다. 3장에서는 본 연구가 제안한 모델의 구조와 학습 방법을 설명하고, 4장에서는 본 연구의 실험 환경과 성능 평가 결과를 정리한다. 마지막으로 5장에서는 연구 내용을 요약하고, 본 연구의 한계점 및 향후 연구에 대해 기술한다.

Ⅱ. 관련 연구

2.1 BART

대표적인 사전학습 언어모델 중 하나인 BART (Bidirectional and Auto-Regressive Transformers)는 트랜스포머(Transformer)[2] 구조를 기반으로 하는 언어모델로서, 언어의 이해와 생성에서 우수한 성능을 발휘한다. BART는 인코더-디코더 구조를 기반으로 잡음(Noise)이 추가된 데이터를 복구하는 방식으로 학습이 진행된다. 인코더는 잡음이 추가된 텍스트를 디코더에 전달하며 디코더는 출력과 원본 텍스트의 손실값을 줄이는 것을 목표로 학습이 진행된다.

노이즈 생성에 사용되는 대표적인 기법으로는 토큰 마스킹(Token masking), 토큰 제거(Token deletion), 텍스트 채우기(Text infilling), 문장 순열(Sentence permutation) 및 문서 회전(Document rotation) 방식이 있다[1]. 한국어를 학습한 BART인 KoBART 모델은 텍스트 채우기 함수를 사용해 40GB 이상의 한국어 텍스트를 학습한 한국어 사전학습 언어모델이다[3].

2.2 도메인 특화 학습

BART 등의 사전학습 언어모델은 범용적인 말뭉치를 기반으로 학습되어 일반적인 도메인에 대한 자연어 처리는 높은 성능을 보이지만 의료나 법률 등과 같은 특수한 도메인에서는 성능이 저하되는 모습을 보인다. 이는 특정 도메인에서만 사용하는 단어로 인한 OOV(Out-of-Vocabulary)와 유의어 간의 관계에 대한 이해 부족으로 발생하는 현상이다.

이처럼 학습에 사용되는 말뭉치에 따라서 언어모

델의 성능이 결정되기 때문에 도메인에 특화된 데이터셋 구축과 학습모델 생성을 위한 연구가 제반 분야에서 활발하게 진행되고 있다. [4]에서는 의료 도메인 말뭉치를 수집하여 이를 BERT(Bidirectional Encoder Representations from Transformers)[5] 모델에 학습시킴으로써 의료 도메인에 특화된 Bio-BERT를 개발하였다. 그 결과 개체명 인식(Named entity recognition), 관계 추출(Relatioin extraction), 질의응답(Question answering) 등에서 기존 BERT 모델을 뛰어넘는 성능을 보였다. [6]에서는 경제 뉴스, 금융상품 설명서 등을 기반으로 구축한 금융 특화 말뭉치를 사용하여 금융 도메인에 특화된 BERT 모델로 학습시켰으며, 금융 관련 지식이 요구되는 자연어 처리에서 비교 대상 모델을 뛰어넘는 성능을 보였다. [7]에서는 법률 판례 분류 기능을 개선하기 위해 각 범죄에 특화된 BERT 모델을 사용하였다.

국방 분야 또한 이러한 도메인 특화 모델의 필요성이 존재하는 분야이다. 일반적으로 사용되지 않는 군사용어나 약어, 각종 체계 명칭 등은 언어모델에 학습을 시켜야 군에서 활용 시 언어모델이 해당 용어를 인식하고 이해할 수 있다. 이와 관련된 연구로 [8]에서는 국방일보와 군사 뉴스로부터 구축한 군사 말뭉치를 통해 군사 도메인에 특화된 BERT 모델을 제안하였으며, 군사 문장 이진분류 실험에서 원본 BERT보다 우수한 성능을 보임을 확인하였다.

2.3 적응형 토크나이저

적응형 토크나이저(Adaptive Tokenizer)는 특정 도메인이나 작업에 맞춰서 언어모델의 토크나이저를 조정하거나 변경하는 기술을 의미한다. 적응형 토크나이저를 사용하면 기존 토크나이저를 수정하거나 확장하여 특정 도메인에서 사용되는 용어나 표현을 언어모델이 올바르게 이해하게 된다. 그리고 미세조정(fine-tuning)과는 달리 전체 모델을 다시 학습하지 않고 특정 도메인에 특화된 언어 처리를 수행할 수 있기 때문에 시간과 계산 리소스를 절약할 수 있다. 이와 관련된 연구로 [9]에서는 적응형 토크나이저를 이용하여 범용 모델의 OOV를 방지하고 특정 도메인에 효율적으로 적응하는 방법을 제안하였다.

III. 군사 도메인 특화 사전학습 언어모델

3.1 Mil-BART

그림 1은 본 연구에서 제안하는 군사 도메인에 특화된 Mil-BART 모델의 전체적인 학습과정을 보여주고 있다. 먼저, 국방 표준 자연어 데이터셋을 구축하고 있는 단독망 환경에서 비공개 자료인 육군·해군·공군 및 해병대 군사교범 데이터셋을 이용하여 군별 군사교범의 토큰을 추출하였다. 이후, BART의 기존 토크나이저에 군사교범 토큰을 추가하여 Mil-BART(비공개 자료) 모델을 구축하였다.

실험은 Window 10 Pro 기반에 10th Gen Intel i9-10980XE CPU와 NVIDIA RTX 4090 2-way GPU, 128GB RAM이 탑재된 데스크탑 환경에서 진행되었고, 사용한 개발언어는 Python 3.10이다.

국방 표준 자연어 데이터셋 중 공개 자료인 국방논단 데이터셋을 이용한 실험도 동일한 방법으로 진행하였다. 추출된 국방논단의 토큰을 기존 BART 토크나이저에 추가하여 새로운 Mil-BART(공개 자료) 모델을 구축한 후 일반 뉴스 문장과의 이진분류 실험에 활용하였다. 국방논단 데이터셋을 이용한 실험은 구글에서 제공하는 Colab pro 환경(Python 3.10, T4 GPU, 51GB RAM)에서 진행되었다.

3.2 군사용어 토크나이저 구축

기존 BART를 군사 도메인에 특화시키기 위해서는 군사용어가 포함된 토크나이저가 필요하다. 기존 BART의 토크나이저는 특정 도메인에 있는 단어를 여러 개의 하위 단어로 분절하여 전문용어가 가진 고유한 의미가 사라질 수 있다. 예를 들어, 군사작전에서 자주 사용되는 용어인 '근접항공지원(CAS)'이라는 용어를 토큰화하면 BART의 토크나이저는 '근접', '항공' 및 '지원'으로 분절하여 학습함으로써 하나의 토큰으로 인식해야 할 '근접항공지원'의 의미가 상실된다. 따라서 군사용어 토크나이저를 구축할 때에는 군사 전문용어가 가진 고유한 의미가 사라지지 않도록 군사용어의 토큰화 과정에서 주의를 기울여야만 한다.

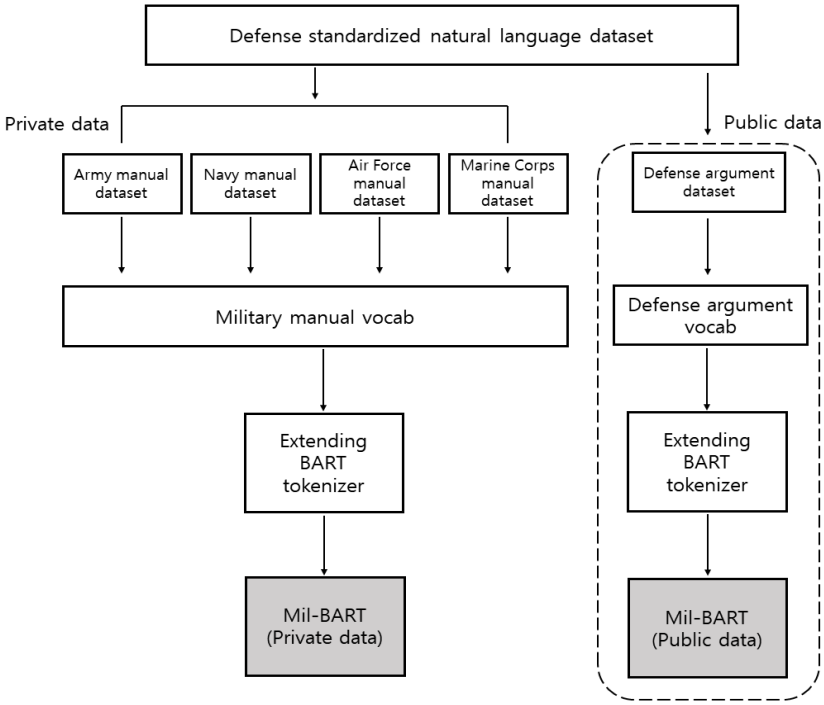

그림 1. Mil-BART의 전체 구조
Fig. 1. Overall architecture of Mil-BART

표 1에서는 본 연구에서 제안하는 Mil-BART의 토크나이저에 새롭게 추가한 토큰의 세부 사항을 확인할 수 있다. 비공개 자료로 실험한 Mil-BART에는 육군·해군·공군 및 해병대 군사교범에 포함되어 있는 토큰 71,253개 중 중복되는 토큰 22,146개를 제외한 49,107개의 토큰을 기존 토크나이저에 추가시켰다. 공개 자료로 실험한 Mil-BART에는 국방논단 데이터셋 중 중복을 제외한 55,350개의 토큰을 기존 토크나이저에 추가시켰다.

표 1. Mil-BART 토크나이저에 추가된 토큰의 세부 사항
Table 1. Details of tokens added to Mil-BART tokenizer

Dataset		Tokens	Ratio
Private data	Army manual	13,883	19.5%
	Navy manual	14,408	20.2%
	Air Force manual	24,997	35.1%
	Marine corps manual	17,965	25.2%
	Total	71,253	100%
Public data	Defense argument	55,350	100%

IV. 실험 및 결과

4.1 군사교범 - 비군사 문장 분류

군사교범과 비군사 문장 간의 이진분류 및 다중분류 실험을 진행하기 위해 표 2에서 보는 바와 같이 각 데이터에 대한 레이블링 작업을 수행하였다.

표 2. 실험용 데이터셋에 대한 레이블 부여 현황
Table 2. Labels of experimental dataset

Dataset	Binary classification labels	Multiclass classification labels	Sentences
Army manual	0	0	4,494
Navy manual	0	1	4,539
Air Force manual	0	2	6,319
Marine Corps manual	0	3	6,377
Genaral news	1	4	8,561
Total	-	-	30,290

비군사 문장은 인터넷을 통해 수집한 일반 뉴스 문장을 활용하였다. 그리고 군사교범 문장은 전부 '0'으로, 일반 뉴스 문장은 '1'로 레이블링하여 이진분류 실험을 진행하였으며, 다중분류 실험을 위해서는 각 군별 교범 문장과 일반 뉴스 문장에 0에서 4까지의 레이블을 각각 부여하였다.

이진분류와 다중분류 실험의 세부 과정은 그림 2에서 보는 바와 같다. 3 epoch 학습을 진행하였고, 분류모델 성능평가에 주로 사용되는 Precision, Recall, F1-score 및 Accuracy를 평가지표로 활용하여 성능을 분석하였다.

모델이 군사 관련 용어에 대한 이해가 높은지 확인하기 위해선 군사 문장을 효과적으로 식별하는 것이 중요한 작업이므로 실제 군사 문장의 샘플 중 모델이 군사 문장으로 정확하게 예측한 비율인 Recall 점수가 중요하며, Precision과 Recall의 조화평균인 F1-score도 중요한 지표이다.

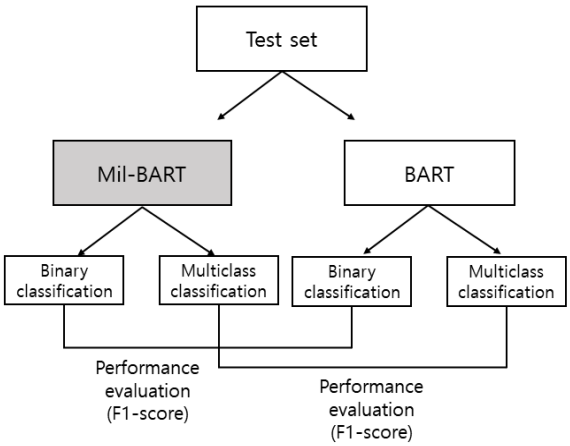

그림 2. 군사교범 - 비군사 문장 분류 실험과정
Fig. 2. Classification process of military manual sentence and non-military sentence

4.2 국방논단 - 비군사 문장 이진분류

국방논단과 일반 뉴스 문장 간 이진분류 실험은 표 3에서 보는 바와 같이 국방논단 문장을 '0'으로 일반 뉴스 문장을 '1'로 레이블링하여 진행하였다.

이진분류 실험과정은 그림 3에서 확인할 수 있으며, 학습은 2epoch를 진행한 후 생성된 Mil-BART를 이용하여 성능을 분석하였다.

표 3. 국방논단 - 일반 뉴스의 이진분류를 위한 레이블
Table 3. Labels for binary classification between military argument sentence and general news sentence

Dataset	Labels	Sentences
Defense argument	0	6,862
Genaral news	1	8,561
Total	-	15,423

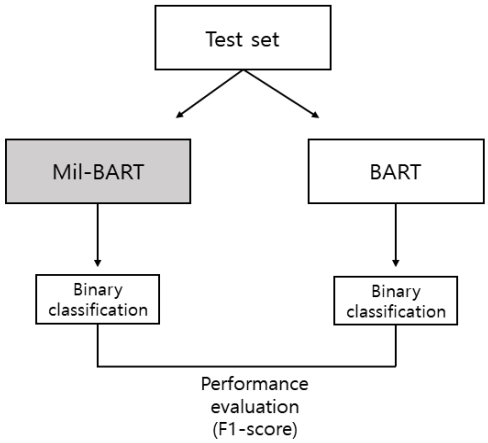

그림 3. 이진분류 실험 과정: 국방논단 - 일반 뉴스
Fig. 3. Experimental process of binary classification: Defense argument-General news

4.3 실험 결과 및 분석

4.3.1 군사교범 - 일반 뉴스 이진분류

군사교범 문장과 일반 뉴스 문장에 대한 이진분류 실험 결과는 표 4에서 보는 바와 같다. 모든 성능 지표면에서 Mil-BART가 BART보다 더 좋은 성능을 보였지만 성능 차이가 크진 않았다.

표 4. 이진분류 성능 평가 결과: 군사교범 - 일반 뉴스
Table 4. Performance evaluation result of binary classification: Military manual-General news

Model	Precision	Recall	F1-score	Accuracy
Mil-BART	0.9953	0.9848	0.99	0.9944
BART	0.9917	0.9772	0.98	0.9913

비교모델인 BART의 경우 토크나이저에 군사용어를 별도로 추가하지 않더라도 군사교범 문장과 비군사 문장의 이진분류에서 탁월한 성능을 보였다.

이러한 결과는 BART와 같은 일반적인 언어모델이 기존 토크나이저에 추가된 일반적인 토큰만으로도 군사교범을 손쉽게 분류할 수 있음을 의미한다. 즉 군사 도메인의 특수성으로 인해 군사교범은 민간에서는 잘 사용되지 않는 특수한 토큰을 다수 포함하고 있어 굳이 군사용어 토큰을 추가하지 않더라도 분류가 용이한 것으로 해석할 수 있다.

군사교범 문장과 일반 뉴스 문장의 이진분류에 대한 혼동행렬은 표 5 및 표 6에서 확인할 수 있다.

표 5. 이진분류 혼동행렬(BART): 군사교범 - 일반 뉴스
Table 5. Confusion matrix of binary classification(BART): Military manual-General news

Class		Predicted labels	
		Military manual	Genaral news
True labels	Military manual	4332	14
	General news	39	1673

표 6. 이진분류 혼동행렬(Mil-BART): 군사교범 - 일반 뉴스
Table 6. Confusion matrix of binary classification(Mil-BART): Military manual-General news

Class		Predicted labels	
		Military manual	Genaral news
True labels	Military manual	4338	8
	General news	26	1686

4.3.2 국방논단 - 일반 뉴스 이진분류

다음으로 국방논단 문장과 일반 뉴스 문장 간의 이진분류 실험 결과는 표 7에서 보는 바와 같다. 군사교범 문장과 일반 뉴스 문장 이진분류 심험에서 확인했던 미미한 성능 향상과는 달리 Mil-BART가 BART보다 모든 성능 지표에서 2%P 정도 성능이 향상되었다.

이는 '군사교범-일반 뉴스 문장 이진분류' 보다 일반 뉴스 문장과 더 비슷한 '국방논단-일반 뉴스 문장 이진분류'가 상대적으로 어렵기 때문에 군사용어를 추가한 토크나이저가 좀 더 향상된 분류성능을 보인 것으로 해석할 수 있다.

표 7. 이진분류 성능 평가 결과: 국방논단-일반 뉴스
Table 7. Performance evaluation result of binary classification: Defense argument-General news

Model	Precision	Recall	F1-score	Accuracy
Mil-BART	0.9625	0.9585	0.96	0.9562
BART	0.9428	0.9433	0.94	0.9368

예를 들어, 일반인에게 전차(군사교범)와 자동차(일반 뉴스)의 분류는 그리 어렵지 않은 작업이다. 그리고 일반인에게 전차가 가지는 다양한 특징을 추가로 학습시킨다 하더라도 분류하는 능력이 크게 향상되지는 않는다. 반면 군에서 운용하는 승용차량(국방논단)을 일반 승용차량(일반 뉴스)과 분류하는 작업은 쉽지 않다. 번호판의 형식이나 기타 차량에 부착된 표식, 탑승자의 복장 등 세부적인 특징들을 정확하게 학습해야만 분류가 가능해진다. 즉 군사 도메인의 특수성이 명확한 전차는 일반인에게 추가적인 학습을 시키더라도 분류성능이 크게 향상되지 않지만, 분류가 어려운 군용 승용차량과 일반 승용차량의 분류는 반드시 추가적인 학습을 실시해야 분류성능이 향상될 수 있다.

국방논단-일반 뉴스 문장에 대한 이진분류 혼동행렬은 표 8과 표 9에서 확인할 수 있다.

표 8. 이진분류 혼동행렬(BART): 국방논단-일반 뉴스
Table 8. Confusion matrix of binary classification(BART): Defense argument-General news

Class		Predicted labels	
		Defense argument	Genaral news
True labels	Defense argument	1275	98
	General news	97	1615

표 9. 이진분류 혼동행렬(Mil-BART): 국방논단-일반 뉴스
Table 9. Confusion matrix of binary classification(Mil-BART): Defense argument-General news

Class		Predicted labels	
		Defense argument	Genaral news
True labels	Defense argument	1309	64
	General news	71	1641

4.3.3 군별 군사교범 - 일반 뉴스 다중분류

각 군별 군사교범 문장과 일반 뉴스 문장에 대한 다중분류 실험 결과는 표 10에서 확인할 수 있으며, 혼동행렬은 표 11 및 표 12에서 보는 바와 같다.

표 10에 나타난 바와 같이 Recall과 Accuracy는 Mil-BART의 성능이 미세하지만 높고, Precision은 낮았다. 그러나 Recall이 미세하게 증가하여 F1-score는 두 모델이 동일하게 나타났다.

Precision의 경우 Mil-BART가 BART보다 성능이 낮게 나타난 이유는 표 11 및 표 12의 혼동행렬에서 확인할 수 있는 바와 같이 Mil-BART가 군별 분류를 수행함에 있어 오분류(FP, False Positive)가 많았기 때문이다. 따라서 군사용어 토크나이저를 가지는 Mil-BART가 군사용어를 좀 더 잘 인식할 수는 있지만 교범의 군별 분류까지 정확하게 수행하지는 못하였다.

표 10. 다중분류 성능 평가 결과: 군사교범 - 일반 뉴스
Table 10. Performance evaluation result of multi-class classification: Military manual-General news

Model	Precision	Recall	F1-score	Accuracy
Mil-BART	0.7766	0.7748	0.77	0.7748
BART	0.7792	0.7745	0.77	0.7745

표 11. 다중분류 혼동행렬(BART): 군사교범 - 일반 뉴스
Table 11. Confusion matrix of multi-class classification(BART): Military manual-General news

Class		Predicted labels				
		Army	Navy	Air force	Marine corps	General news
True labels	Army	715	39	63	75	7
	Navy	140	492	158	114	4
	Air force	94	122	1007	41	0
	Marine corps	217	175	95	783	5
	General news	5	0	0	12	1695

표 12. 다중분류 혼동행렬(Mil-BART): 군사교범 - 일반 뉴스
Table. 12. Confusion matrix of multi-class classification(Mil-BART): Military manual-General news

Class		Predicted labels				
		Army	Navy	Air force	Marine corps	General news
True labels	Army	689	45	64	89	12
	Navy	129	480	151	141	7
	Air force	95	121	1004	44	0
	Marine corps	218	137	87	823	10
	General news	3	0	0	11	1698

표 12의 혼동행렬을 이용하여 Mil-BART의 각 군별 군사교범에 대한 분류성능을 분석한 결과는 표 13에서 보는 바와 같다. 일반 뉴스 문장을 제외하면 공군 교범의 분류성능이 가장 좋았다. 이는 공군에서 사용하는 군사용어들이 타군 대비 범용적이지 않고 전문용어를 더 많이 사용한다는 의미로 해석할 수 있다. 또한, 해병대 교범의 경우 육군 교범 및 해군 교범과 유사한 문장이 많이 포함되어 있어 육군, 해군 및 해병대 교범 간의 분류성능은 상대적으로 낮게 나타난 것으로 보인다. 이를 통해 공군 교범의 군사 도메인 내 특수성이 가장 높다는 것을 확인할 수 있다.

표 13. 군별 군사교범 다중분류 성능 평가 결과 (Mil-BART)
Table 13. Performance evaluation result of multi-class classification(Mil-BART): Classification by military service

Class	Precision	Recall	F1-score	Accuracy
Army	0.8775	0.7664	0.8182	0.8775
Navy	0.8653	0.5286	0.6563	0.8653
Air force	0.8931	0.7943	0.8408	0.8931
Marine corps	0.8643	0.6455	0.7390	0.8643
General news	0.9909	0.9918	0.9914	0.9909

V. 결론 및 향후 과제

본 논문에서는 현재 구축사업이 진행되고 있는 국방 표준 자연어 데이터셋을 최초로 적용해 군사 도메인에 특화된 사전학습 언어모델인 Mil-BART 모델을 제안하고, 군사용 문장(군사교범 및 국방논단)과 비군사용 문장(일반 뉴스)에 대한 이진분류 및 다중분류 실험을 진행한 후 성능을 분석하였다.

그 결과 이진분류에서는 Mil-BART가 모든 성능 지표에서 BART보다 뛰어난 것으로 나타났고, 특히 국방논단 문장과 일반 뉴스 문장에 대한 이진분류 성능은 2%P 정도 향상됨을 확인하였다. 그러나 각 군별 군사교범 분류를 시도한 다중분류 실험에서는 공군 군사교범을 제외하고는 다수의 오분류가 확인되어 군사용어를 추가한 언어모델이어도 모든 분류에서 성능이 향상되는 것은 아니라는 점도 알 수 있었다. 따라서 민간에서 활용되고 있는 언어모델을 군사분야에 적용하기 위해서는 표준화된 데이터셋 구축뿐만 아니라 자연어처리 기법과 관련된 연구도 적극적으로 실시해야 할 것으로 판단된다.

민간에서 개발된 언어모델은 날이 갈수록 성능이 향상되고 있으며, 굳이 군사용어를 학습하지 않은 상태에서도 군사교범 문장을 비교적 확실하게 분류하였다. 그러나 국방논단-일반 뉴스 문장 분류에서 확인된 바와 같이 그 특수성이 명확하게 드러나지 않는 문장들에 대한 분류 성능은 떨어진다. 따라서 군사용어를 완벽하게 이해하고 문장 분석도 효과적으로 수행할 수 있는 군사 도메인에 특화된 언어모델의 개발은 반드시 필요하다. 그러한 측면에서 본 연구는 과학적 근거를 통해 군사 도메인에 특화된 언어모델 개발의 중요성을 입증하고, 현재 구축을 진행 중인 국방 표준 자연어 데이터셋이 군 내 AI 학습용 데이터로 충분히 활용 가능함을 검증한 최초의 연구라고 할 수 있다.

향후에는 국방 표준 자연어 데이터셋의 일부로서 함께 구축되고 있는 음성 데이터셋을 활용하면서 자연어처리와 연계된 연구를 지속해 나갈 예정이다.

References

[1] M. Lewis, et al., "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension", Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871-7880, Jul. 2020. https://doi.org/10.18653/v1/2020.acl-main.703.

[2] A. Vaswani, et al., "Attention is All You Need", Advances in Neural Information Processing Systems 30, pp. 6000-6010, Dec. 2017.

[3] Github, "KoBART", last modified, May 2022. https://github.com/SKT-AI/KoBART [accessed: Oct. 25, 2023]

[4] J. Lee. W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", Bioinformatics, Vol. 36, No. 4, pp. 1234-1240, Feb. 2020. https://doi.org/10.1093/bioinformatics/btz682.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, Vol. 1, pp. 4171-4186, Jun. 2019. http://dx.doi.org/10.18653/v1/N19-1423.

[6] D. Kim, D. Lee, J. Park, S. Oh, S. Kwon, I. Lee, and D. Choi, "KB-BERT: Training and Application of Korean Pre-trained Language Model in Financial Domain", Journal of Intelligence and Information Systems, Vol. 28, No. 2, pp. 191-206, Jun. 2022.

[7] S. Han, B.-W. On, and D. Jeong, "Research on BERT and Similarity-Based Models for Legal Precedent Search", Proc. of KIIT Conference, Jeju, Korea, pp. 585-589, Nov. 2023.

[8] H.-S. Heo, C.-M. Yoon, Y.-H. Ryu, S. Yong, and D. Kim, "MIL-BERT: Military Domain Specialized

Korean Pre-trained Language Model", Journal of the KNST, Vol. 6, No. 2, pp. 201-206, Jun. 2023. https://doi.org/10.31818/JKNST.2023.06.6.2.201.

[9] V. Sachidananda. J. Kessler, and Y.-A. Lai, "Efficient domain adaptation of language models via adaptive tokenization", Proc. of the Second Workshop on Simple and Efficient Natural Language Processing, pp. 155-165, Nov. 2021. https://doi.org/10.18653/v1/2021.sustainlp-1.16.

저자소개

고 관 우 (Gwan-Woo Goh)

2016년 2월 : 육군사관학교
토목환경공학과(공학사)
2024년 1월 : 국방대학교
국방과학학과(국방과학석사)
관심분야 : 자연어처리, 인공지능

김 한 석 (Han-Seok Kim)

2014년 2월 : 육군사관학교
국제관계학과(문학사)
2024년 1월 : 국방대학교
국방과학학과(국방과학석사)
관심분야 : 머신러닝, 침입 탐지 시스템, 자연어처리, 인공지능

이 수 진 (Soo-Jin Lee)

1992년 : 육군사관학교
컴퓨터공학과(공학사)
1996년 : 연세대학교
컴퓨터공학과(공학석사)
2006년 : 한국과학기술원(KAIST)
컴퓨터공학과(공학박사)
2006년 ~ 현재 : 국방대학교
국방과학학과 교수
관심분야 : 국가 사이버 보안 정책, 침입 탐지 시스템, 모바일 네트워크 보안, 머신 러닝, 암호화 이론 및 응용