KorBERT 기반 빈칸채우기 문제를 이용한 텍스트 분류

허정○, 이형직, 임준호

한국전자통신연구원, 언어지능연구실
Jeonghur, leehj, joonho.lim@etri.re.kr

Text Classification using Cloze Question based on KorBERT

Jeong Heo○, Hyung-Jik Lee, Joon-Ho Lim
ETRI, Language Intelligence Research

요 약

본 논문에서는 KorBERT 한국어 언어모델에 기반하여 텍스트 분류문제를 빈칸채우기 문제로 변환하고 빈칸에 적합한 어휘를 예측하는 방식의 프롬프트기반 분류모델에 대해서 소개한다. [CLS] 토큰을 이용한 헤드기반 분류와 프롬프트기반 분류는 사전학습의 NSP모델과 MLM모델의 특성을 반영한 것으로, 텍스트의 의미/구조적 분석과 의미적 추론으로 구분되는 텍스트 분류 태스크에서의 성능을 비교 평가하였다. 의미/구조적 분석 실험을 위해 KLUE의 의미유사도와 토픽분류 데이터셋을 이용하였고, 의미적 추론 실험을 위해서 KLUE의 자연어추론 데이터셋을 이용하였다. 실험을 통해, MLM모델의 특성을 반영한 프롬프트기반 텍스트 분류에서는 의미유사도와 토픽분류 태스크에서 우수한 성능을 보였고, NSP모델의 특성을 반영한 헤드기반 텍스트 분류에서는 자연어추론 태스크에서 우수한 성능을 보였다.

주제어: KorBERT, 텍스트분류, 프롬프트기반 분류, 헤드기반 분류

1. 서론

딥러닝 기술의 발전으로 대용량 텍스트 데이터를 이용한 자기지도(self-supervised) 학습으로 자연어의 의미 표현을 사전학습(pre-training)하고 이를 이용하여 다양한 응응 태스크에 활용하는 언어모델(language model)이 많이 연구되고 있다. 특히, 트랜스포머(transformer)의 인코더(encoder)를 이용한 BERT계열의 언어이해 모델은 다양한 태스크에서 우수한 성능을 보이고 있다[1,2,3].

BERT는 주변 단어를 이용하여 랜덤하게 마스킹(masking)된 단어를 예측하는 MLM 모델(masked language model)과 두개의 문장이 문맥적으로 이어지는 문장인지를 예측하는 NSP(next sentence prediction)모델로 사전학습이 수행된다. MLM모델은 자연어의 문장 의미 및 구조적 특성을 학습하고, NSP는 자연어에 대한 문장간 문맥 의미를 이해하도록 학습된다[1].

BERT를 이용한 분류문제는 일반적으로 [CLS] 토큰을 입력으로 받는 헤드 레이어(head layer)를 이용하여 텍스트 분류(헤드기반 텍스트 분류)를 한다. 그러나, 분류문제를 태스크에 대한 설명과 함께 빈칸채우기 문제로 변환하고, 해당 빈칸에 대한 단어를 예측하도록 하여, 사전학습의 MLM모델을 학습하는 방식으로 해결하는 연구(프롬프트기반 텍스트분류)가 진행되었다[4]. 헤드기반의 텍스트분류는 사전학습에서 NSP의 특성을 반영하고, 프롬프트기반 텍스트분류는 MLM의 특성을 반영한다.

텍스트 분류 태스크는 문장의 의미/구조적 분류와 문장간 의미추론으로 구분될 수 있다. 문장의 의미/구조적 분류는 문장 의미유사도 측정, 문장의 토픽 분류 등이 있고, 문장간 의미추론은 문장의 함의관계를 추론하는 태스크 등이 있다. 헤드기반 텍스트 분류와 프롬프트기반 텍스트 분류는 사전학습의 NSP모델과 MLM모델의 특성을 각자 반영하고 있으므로, 언급한 두종류의 텍스트 분류 태스크에서 다른 특성을 보일 것으로 추정된다.

본 논문에서는 KorBERT[1]를 이용하여 분류문제를 빈칸채우기 문제로 변환하여 예측하는 프롬프트 기반 텍스트 분류모델에 대해서 소개하고, 헤드기반 텍스트 분류모델과 다양한 한국어 분류 태스크에서 비교실험을 하였다. 문장간 의미추론 태스크와 문장의 의미/구조적 분류 태스크 간의 실험을 통해 프롬프트 기반 텍스트 분류 모델이 문장의 의미/구조분석 태스크에 더울 효과적임을 확인하였다.

2. 관련 연구

PET(pattern exploiting training) 알고리즘은 PVPs(pattern verbalizer pairs)를 이용하여 빈칸채우기 형태로 변환하여, 언어모델의 사전학습 방식을 크게 벗어나지 않게 하면서 세부 태스크(downstream task)에 대해서 퓨샷 학습(few-shot learning) 성능을 개선시키는 방법을 소개한다. 또한, PET 알고리즘을 이용하여 라벨링되지 않은 데이터를 라벨링하여 학습데이터로 만들어 다시 모델을 학습하는 과정을 몇 세대(generation) 반복하

[1] https://aiopen.etri.re.kr/service_dataset.php

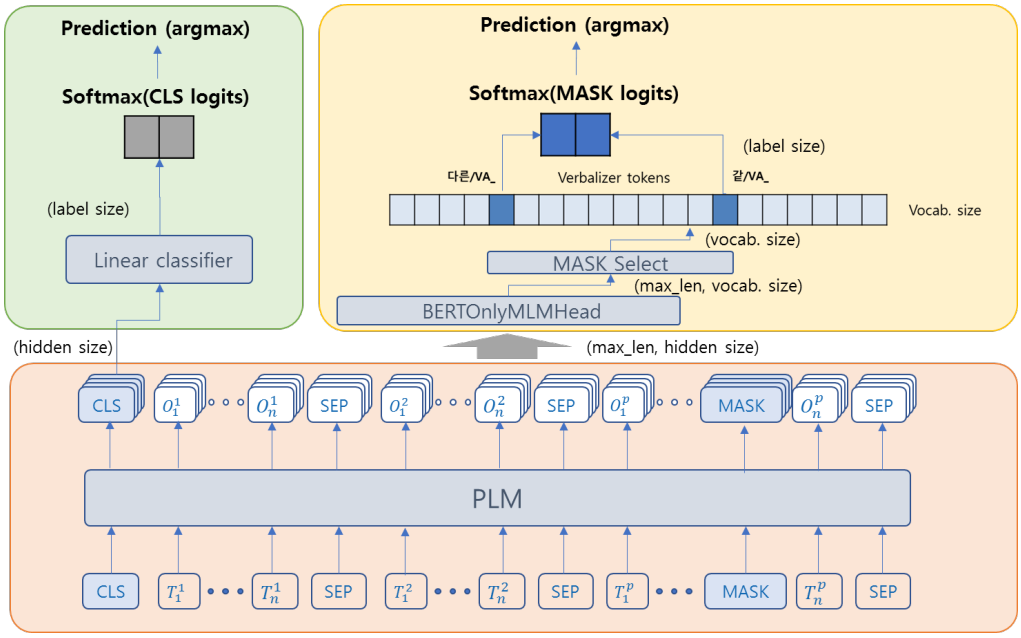

그림 1. KorBERT를 이용한 헤드기반 분류와 프롬프트기반 분류

여 성능을 개선한 iPET 알고리즘도 제안하고 있다[4].

[5]에서는 텍스트 분류를 위한 전이학습(transfer learning)의 두가지 방법인 헤드 레이어를 이용하는 헤드기반 전이학습과 태스크에 특화된 패턴 문장을 이용하는 프롬프트기반 전이학습를 비교분석하고 있다. 논문에서는 프롬프트 기반의 전이학습 방법이 헤드기반 전이학습보다 효율성 측면에서 상당한 이점이 있음을 실험을 통해 보이고 있다.

본 논문에서는 KorBERT를 이용하여 한국어 텍스트 분류문제에서 빈칸채우기 형태의 프롬프트 설계하고, 이를 이용하여 KLUE 벤치마크 데이터(문장의미/구조 분석, 문장간 의미추론)를 대상으로 실험을 수행하여, 헤드기반 텍스트 분류모델과 프롬프트기반 텍스트 분류모델의 성능을 비교하였다.

3. 빈칸채우기 문제를 이용한 분류모델(CQC 모델)

[그림 1]은 [CLS]를 이용한 헤드기반 분류와 프롬프트기반 분류를 그림으로 비교 제시하고 있다. [그림 1]은 두 개의 텍스트($T_1^1 \sim T_n^1$과 $T_1^2 \sim T_n^2$)에 대한 의미적 유사성을 분류하는 예시로써, 헤드기반 분류에서는 [CLS + $T_1^1 \sim T_n^1$ + SEP + $T_1^2 \sim T_n^2$ + SEP]이 입력되고, [CLS] 토큰에 대해 이진분류를 하는 linear 레이어를 거쳐 softmax하여 분류한다. 반면, 프롬프트기반 분류에서는 마스킹된 토큰을 가지는 프롬프트($T_1^p \sim T_n^p$)를 두 개의 텍스트와 연결한 [CLS + $T_1^1 \sim T_n^1$ + SEP + $T_1^2 \sim T_n^2$ + SEP + $T_1^p \sim T_n^p$ + SEP]을 입력하고, 프롬프트 내에 마스킹된 토큰에 대해 미리 정의한 클래스와 매핑되는 어휘토큰들에 대한 softmax를 통해 분류를 한다. 이를 위해서는 프롬프트로 입력되는 패턴(pattern)을 사전에 입력해야 하고, 마스킹된 토큰과 분류한 클래스별로 매핑될 어휘사전의 토큰(verbalizer)을 정의해야 한다. 패턴과 마스킹된 토큰에 대한 어휘사전 토큰 목록을 PVPs(pattern verbalizer pairs)라고 한다.

4. KLUE 벤치마크와 PVPS

본 논문에서는 한국어에 대한 자연어 이해 벤치마크 데이터 셋인 KLUE(Korean language understanding evaluation benchmark)에서 텍스트 분류와 관련된 아래의 세가지 데이터 셋을 이용하였다[6].

- 문장 의미/구조 분석 태스크
  - 의미유사도(semantic textual similarity): 주어진 두 문장 간의 의미 동등성을 수치로 표현하는 문제.(본 논문에서는 이진분류 레이블을 이용한 분류문제로 평가함)
  - 토픽분류(topic classification): 뉴스 헤드라인에 대해 7가지 클래스(정치, 경제, 사회, 생활문화, 세계, IT과학, 스포츠)로 분류하는 문제
- 문장간 의미추론 태스크
  - 자연어추론(natural language inference): 전제로 주어진 텍스트와 가설로 주어진 텍스트 간의 관계를 추론하는 문제(entailment, contradiction, neutral로 분류하는 문제)

KLUE의 세가지 데이터 셋에 대한 PVPs는 [표 2]~[표 4]와 같이 정의하였다. 각 PVPs에서 random은 모델이 정의된 패턴들 중에 무작위로 선택하여 학습하도록 하는 패턴이다. 태스크 별 프롬프트과 verbalizer는 경험적으로 선택하였다. 마스킹된 토큰은 [MASK]로 표시하였고, verbalizer는 '분류코드:예측토큰'으로 구성하였다. pattern은 해당 프롬프트가 분류하려는 텍스트(들)중

어느 위치에 입력될 것인지를 결정한다. 프롬프트의 위치는 프롬프트가 문맥적으로 자연스러운 위치를 대상으로 선정하였다. 예를 들면, 의미유사도 PVP 0번의 경우. 프롬프트가 “앞뒤/NNG 문장/NNG 은/JX [MASK] 다/EF ./SF”이므로, 프롬프트의 위치는 두 텍스트의 중간에 위치하는 것이 문맥상 자연스럽다. 그리고, verbalizer는 프롬프트에서 마스킹된 위치에 들어갈 수 있는 어휘토큰으로 분류하고자 하는 클래스와 의미적으로 부합하는 어휘토큰을 선정하였고, 프롬프트 문장과도 의미적으로 완전한 문장이 될 수 있는 어휘토큰이다.

5. 실험 및 평가

[표 1]은 실험에 사용된 KLUE 데이터의 통계 정보이다. 테스트 데이터는 공개되지 않으므로 개발데이터로 평가를 수행하였다.

표 1. KLUE 데이터 셋 통계 정보

구분	의미유사도 (STS)	토픽분류 (TC)	자연어추론 (NLI)
학습데이터	11,668	45,678	24,998
개발데이터	519	9,107	3,000

실험은 KorBERT_base 언어모델을 이용하였고, batch 사이즈는 16, max_epoch는 5, learning rate는 2e-5, max_length는 128로 설정하였다.

실험 중 가장 높은 성능을 보인 모델로 헤드기반 분류와 프롬프트기반 분류를 비교하였다. 의미유사도는 평가척도로 F1, 자연어추론은 정확률(accuracy), 토픽분류는 macro-F1을 사용하였다.

표 5. 평가결과

태스크 (metric)	헤드기반 분류	프롬프트 기반 분류	
	Best score		Best pattern
STS (F1)	0.8404	**0.8442**	0 번 PVP
TC (F1)	0.8651	**0.8710**	1 번 PVP
NLI (Acc.)	**0.8137**	0.7993	2 번 PVP

[표 5]는 평가결과로서, 문장의 의미/구조적 분류에 해당하는 문장유사도(STS)와 토픽분류(TC) 태스크에서는 MLM모델의 특성을 반영한 프롬프트기반의 텍스트 분류 모델이 0.38%와 0.59%의 성능 우수를 보였다. 반면, 문장간 의미 추론이 필요한 자연어추론(NLI) 태스크에서는 NSP모델의 특성을 반영한 헤드기반의 텍스트 분류 모델이 1.44%의 성능 우수를 보였다.

MLM모델의 특성을 이용한 프롬프트기반 텍스트 분류

표 2. KLUE 의미유사도(STS) PVPs

#	prompt	verbalizer	pattern
0	앞뒤/NNG 문장/NNG 은/JX [MASK] 다/EF ./SF	{ 0: "다르/VA_", 1: "같/VA_" }	[CLS]sent1[SEP]**prompt**[SEP]sent2[SEP]
1	두/MM 문장/NNG 의/JKG 의미/NNG 가/JKS [MASK] 다/EF ./SF		[CLS]**prompt**[SEP]sent1[SEP]sent2[SEP]
2	두/MM 문장/NNG 의/JKG 의미/NNG 가/JKS [MASK] 다/EF ./SF		[CLS]sent1[SEP]sent2[SEP]**prompt**[SEP]
3	[MASK] 다/EF ./SF		[CLS]sent1[SEP]sent2[SEP]**prompt**[SEP]
4	[MASK] 의미/NNG 이/VCP 다/EF ./SF	{0: "다른/MM_", 1: "동일/NNG_"}	[CLS]sent1[SEP]sent2[SEP]**prompt**[SEP]
random	0~4의 패턴을 랜덤하게 선택하여 학습		

표 3. KLUE 토픽분류(TC) PVPs

#	prompt	verbalizer	pattern
0	주제/NNG 는/JX [MASK] 이/VCP 다/EF ./SF	{ 0:"정치/NNG_", 1:"경제/NNG_", 2:"사회/NNG_", 3:"문화/NNG_", 4:"세계/NNG_", 5:"과학/NNG_", 6:"스포츠/NNG " }	[CLS]**prompt**[SEP]sent[SEP]
1	주제/NNG 는/JX [MASK] 이/VCP 다/EF ./SF		[CLS]sent[SEP]**prompt**[SEP]
random	0~1의 패턴을 랜덤하게 선택하여 학습		

표 4. KLUE 자연어추론(NLI) PVPs

#	prompt	verbalizer	pattern
0	두/MM 문장/NNG 은/JX [MASK] 관계/NNG 이/VCP 다/EF ./SF	{ 0: "동의/NNG_", 1: "중립/NNG_", 2: "반대/NNG_" }	[CLS]sent1[SEP]**prompt**[SEP]sent2[SEP]
1	두/MM 문장/NNG 은/JX [MASK] 관계/NNG 이/VCP 다/EF ./SF		[CLS]**prompt**[SEP]sent1[SEP]sent2[SEP]
2	두/MM 문장/NNG 은/JX [MASK] 관계/NNG 이/VCP 다/EF ./SF		[CLS]sent1[SEP]sent2[SEP]**prompt**[SEP]
3	[MASK] 이/VCP 다/EF ./SF		[CLS]sent1[SEP]sent2[SEP]**prompt**[SEP]
4	함의/NNG 관계/NNG [MASK] 다/EF ./SF	{ 0: "맞/VV_", 1: "모르/VV_", 2: "아니/VCN_" }	[CLS]sent1[SEP]sent2[SEP]**prompt**[SEP]
random	0~4의 패턴을 랜덤하게 선택하여 학습		

는 문장의 의미/구조 분석 태스크에 적합하고, NSP모델의 특성을 이용한 헤드기반 텍스트 분류는 문장간 의미 추론 태스크에 적합한 것으로 판단된다.

5. 결론

본 논문에서는 KorBERT를 이용하여 분류문제를 빈칸채우기 문제로 변환하여 예측하는 프롬프트기반 텍스트 분류모델에 대해서 소개하였고, [CLS] 토큰을 이용한 헤드기반 분류모델과 프롬프트기반 분류모델을 문장의 의미분류 태스크와 문장간 의미추론 태스크로 구분하여 분류모델별 특성을 분석하였다. 사전학습의 MLM 모델을 이용한 프롬프트기반 텍스트 분류 모델에서는 문장의 의미/구조적 분류에 해당하는 문장 유사도와 토픽분류 태스크에서 우수한 성능을 보였고, NSP 모델을 이용하는 헤더기반 텍스트 분류 모델은 문장간 의미적 추론이 필요한 자연어 추론 태스크에서 우수한 성능을 보였다. 본 논문에서는 텍스트 분류 태스크의 특성에 따라, 헤드기반 분류모델과 프롬프트기반 분류모델의 차이를 실험으로 보였으며, 분류모델의 특성이 사전학습의 MLM모델과 NSP모델의 특성에 기인한 것으로 분석하였다.

향후, 프롬프트 기반 텍스트 분류에서 사용되는 프롬프트의 패턴인 PVPs의 구성에 따른 성능 차이를 연구하고, 최적의 PVPs을 선택하기 위한 방법을 연구할 예정이다. 또한, 헤드기반 텍스트 분류모델과 프롬프트기반 텍스트 분류 모델의 하이브리드를 통해, 의미/구조적 특성에 기반한 텍스트 분류와 의미적 추론이 필요한 텍스트 분류에 모두 우수한 성능을 보일 수 있는 모델을 개발할 것이다.

* 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 [No. 2013-2-00131, (엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발].

참고문헌

[1] Jacob Devlin, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAALC, 2019.
[2] Yinhan Liu, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692, 2019
[3] Kevin Clark, et al., ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, arxiv preprint arXiv:2003.10555, 2020
[4] Timo Schick, et al., Exploiting Cloze Questions for F ew Shot Text Classification and Natural Language Inference, arXiv preprint arXiv:2001.07676, 2020
[5] Teven Le Scao, et al., How Many Data Points is a Prompt Worth?, NAACL, 2021
[6] Sungjoon Park, et al., KLUE: Korean Language Understanding Evaluation, arXiv preprint arXiv:2105.09680