상대적 위치 표현을 이용한 한국어 BERT 학습 방법

오연택○, 전창욱, 민경구

LG사이언스파크

yeontaek.oh@lgsp.co.kr, changwook.jun@lgsp.co.kr, kyungkoo.min@lgsp.co.kr

Korean BERT Learning Method with Relative Position Representation

Yeon-Taek Oh○, Chang-Wook Jun, Kyung-Koo Min

LG Sciencepark

요 약

BERT는 자연어처리 여러 응용 분야(task)에서 우수한 성능을 보여줬으나, BERT 사전학습 모델을 학습하기 위해서는 많은 학습 시간과 학습 자원이 요구된다. 본 논문에서는 빠른 학습을 위한 한국어 BERT 학습 방법을 제안한다. 본 논문에서는 다음과 같은 세 가지 학습 방법을 적용했다. 교착어인 한국어 특성을 반영하기 위해 형태소 분석 기반의 사전을 사용하였으며, 단어 간 상대적 위치 표현을 추가하여, 상대적 위치 정보를 학습했다. 또한 BERT 베이스 모델의 12-레이어 중 3-레이어만을 사용하여, 모델을 경량화시켰다.

주제어: BERT, 3-Layer, Relation

1. 서론

BERT(Bidirectional Encoder Representations from Transformer)[1]는 구글에서 발표한 사전학습 모델이다. 다양한 NLP Task에서 SOTA(state-of-the-art)의 성능을 보여주며, 자연어처리 분야의 발전에 크게 기여하였다. 구글에서는 영어와 중국어의 단일 언어 BERT 모델과 104개 언어를 학습한 BERT 다국어 모델을 공개했으며, 최근에는 러시아어 BERT 모델인 RuBERT[2]와 한국어 BERT 모델 KoBERT[3]가 공개되기도 했다. BERT는 자연어처리 여러 응용 분야(task)에서 우수한 성능을 보여줬으나, BERT 사전학습 모델을 학습하기 위해서는 많은 학습 시간과 학습 자원이 요구된다.

본 논문에서는 빠른 학습을 위한 한국어 BERT 학습 방법을 제안한다. 본 논문에서는 세 가지 학습 방법을 적용했다. 첫째, 교착어인 한국어 특성을 반영하기 위해 형태소 분석을 이용해 토크나이징을 진행했다. 둘째, 단어 간 상대적 위치 표현을 추가하여 상대적 위치 정보를 학습했다. 셋째, BERT 베이스 모델의 12-레이어 중 3-레이어만을 사용하여 모델을 경량화시켰다.

본 논문의 구성은 다음과 같다. 2절에서는 BERT 모델과 관련된 연구를 소개하며, 3절에서는 한국어 BERT 모델 학습 방법을 소개한다. 4절에서는 3절에서 제안한 사전학습 모델을 적용한 실험을 기술하며, 5절에서는 논문의 내용을 요약하고 향후 연구에 관해 설명한다.

2. 관련 연구

BERT 모델은 WPM(WordPiece Model)[4] 을 사용하여 부분 단어(sub-word) 학습을 통해 사전을 구성한다. Sub-word 모델은 자주 등장하지 않는 단어는 sub-word unit으로 나눠 학습하고 자주 등장하는 단어는 하나의 unit으로 학습한다. 이를 통해 언어적 지식 없이도 문장의 토크나이징(Tokenizing)이 가능하였으며, BERT 모델의 사전 크기를 줄이고 OOV(Out-of-Vocabulary) 문제를 해결할 수 있었다.

BERT는 NSP(Next Sentence Pre-diction), MLM(Masked Language Modeling)의 2가지 Task로 학습된다.

NSP Task는 두 개의 A, B 문장을 구성하는데, B 문장을 50% 확률로 A 문장의 다음 문장으로 구성하고, 나머지 50%의 확률로 학습 데이터 내 임의의 문장으로 구성한다. 이를 통해 입력된 두 문장이 이어진 문장인지 아닌지를 판별하며, 학습이 진행된다. NSP Task의 목적은 두 문장 사이의 관계를 이해하는데 있으며, Question Answering, Natural Language Inference와 같은 Task에서는 문장 간의 관계를 이해하는 것이 매우 중요하다. NSP Task 유무에 따라 QNLI, MNLI, SQUAD 등의 Task에서 성능 차이를 보여주었다.[1]

BERT 모델은 양방향 학습을 위해 특정 확률만큼 입력 토큰을 마스킹 처리하고 마스킹 된 토큰들을 예측하면서 학습을 진행하는데 이러한 방법을 MLM(Masked Language Modeling)이라고 한다. [1]에서는 입력 문장의 15%를 Random 하게 마스킹 처리했다. 마스킹 대상의 80%는 '[MASK]' 토큰으로 대체하며, 10%는 사전 내 임의의 토큰으로 대체하고 나머지 10%는 원래의 단어를 그대로 사용한다.

최근 이러한 BERT 모델의 학습 방법을 확장시킨 다양한 사전 학습 모델들이 등장했다. 중국어 사전학습 모델인 ERINE[5]는 개체명과 구 단위 기반의 마스킹 전략을 통해 기존 BERT 모델보다 더 나은 성능을 보여주었다. 이는 의미 있는 부분들을 마스킹함으로써 사전 학습모델이 보다 의미론적 표현을 학습할 수 있게 되었기 때문이다. 이와 유사하게 구글에서는 sub-word로 나눠진 토큰 전체를 마스킹 하는 WWM(Whole Word Masking) 모델을 공개하기도 했다.

RoBERTa[6] 모델은 NSP Task를 삭제 후 모델에 문장이 입력될 때 마스킹하는 동적 마스킹(Dynamic Masking)을 적용했다. 이를 통해 다양한 마스킹 토큰들을 학습하는 방법을 보여주었다.

spanBERT[7]는 연속적인 단어를 임의로 마스킹하는 방법을 적용했다. 또한, NSP Task를 삭제하는 대신 마스킹 된 단어의 경계에 있는 단어를 이용해 마스킹 단어를 예측하는 SBO(Span Boundary Objective) Task를 추가하여 학습했다.

본 논문에서는 BERT 모델에서 확장된 한국어 BERT 모델을 제안한다. 3.1에서는 사전 구성 방법을 설명하고 3.2에서는 Relation-aware Self-Attention에 관해 설명한다. 3.3에서는 모델 학습 방법에 관해서 설명한다.

3. 한국어 BERT 모델

3.1 사전 구성

한국어는 하나의 어절이 어근과 접사로 이루어진 교착어이다. [4]와 같은 부분 단어(sub-word) 모델을 사용해 토크나이징 할 경우 사전(Dictionary) 구성에 형태론적 특성을 반영하지 못하여, 모델 전체 성능을 저하하는 원인이 되었다.[8]

본 논문에서는 이러한 문제를 해결하고자 MECAB[9] 한국어 형태소 분석기를 이용해 토크나이징 하고, 형태소 분석 결과를 기반으로 사전을 구성했다. 그러나 형태소 분석 결과를 그대로 이용해 사전을 구성할 경우 사전 크기가 무한으로 늘어나는 문제가 있기 때문에, 형태소 분석 결과의 출현 빈도를 기준으로 사전의 크기를 128,000개 단어로 제한하여 사전을 구축했다.

3.2 Relation-aware Self-Attention

[1]에서는 Transformer[10] 인코더에서 사용했던 Scaled Dot-product Attention으로 모델링 된다. Query의 토큰 x_i와 Key의 토큰 x_j에 각각 가중치 W^Q, W^K를 곱한다. 이렇게 구한 결괏값에 소프트맥스를 취하는데 그 값을 a_{ij}라고 하며, 수식은 아래 (1)과 같다.

$$a_{ij} = Softmax\left(\frac{x_i W^Q (x_j W^K)^T}{\sqrt{d_k}}\right) \quad (1)$$

(1)에서 구한 a_{ij}와 Value 내적합을 통해 결과값 z_i를 구하는데, 수식은 아래 (2)와 같다.

$$z_i = \sum_{j=1}^{n} a_{ij}(x_j W^V) \quad (2)$$

이때 Query, Key, Value는 같은 문장으로 하며, 이러한 Attention Mechanism을 Self-Attention이라고 한다. 전체 수식은 (3)과 같다.

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

[11]에서는 Self-Attention에서 확장된 Relation-aware Self-Attention을 Transformer 모델에 적용하여, 기계번역 Task에서 기존 Transformer 모델보다 높은 성능을 보여주었다.

Relation-aware Self-Attention은 단어 간 상대적 위치 표현을 나타내는 임베딩 테이블을 만들고 Self-Attention 계산 과정에 추가한 것으로 [11]에서는 Query의 토큰 x_i와 Key, Value의 토큰 x_j가 가지는 상대적 위치 임베딩 벡터를 각각 a_{ij}^K, a_{ij}^V로 표현하며, 이를 (1), (2)의 Self-Attention 계산 과정에서 추가했다. 수식은 아래 (4), (5)과 같다.

$$a_{ij} = Softmax\left(\frac{x_i W^Q (x_j W^K + a_{ij}^K)^T}{\sqrt{d_k}}\right) \quad (4)$$

$$z_i = \sum_{j=1}^{n} a_{ij}(x_j W^V + a_{ij}^V) \quad (5)$$

본 논문에서는 BERT 모델의 Position embedding을 삭제하고 Relation-aware Self-Attention를 적용하여, 단어 간 상대적 위치 정보를 이용해 위치(position)를 학습했다. 이를 위해 Query와 Key의 상대적 위치 정보를 Self-Attention 계산 과정에 추가했다.

3.3 학습

본 논문에서는 위키피디아 350만 문장을 사용하여, 128 길이의 문장으로 구성 후 30만 step 학습했다. 모델 레이어는 BERT 베이스 모델의 12-레이어 중 3-레이어만을 사용하였으며, TPU 16장을 이용해 학습을 진행했다. 표 1은 본 논문에서 제안한 모델과 BERT 베이스 모델의 학습 파라미터 비교한 결과이다.

표 1. 모델 학습 파라미터 비교

	our model	BERT, base
레이어	**3**	**12**
히든 사이즈	768	768
헤드 개수	12	12
학습 step	**30만**	**100만**
배치 사이즈	256	256
학습률	1e-4	1e-4

4. 실험 및 결과

본 논문에서는 성능 평가를 한국어 기계 독해 데이터인 KorQuAD 1.0[12]과 네이버 영화 리뷰 감성 말뭉치 NSMC 1.0[13]을 이용해 모델의 성능을 평가했다.

4.1 사전구성 실험 및 결과

표 2는 3.1절에서 제안했던 사전 구성에 관련된 실험 결과이다. BERT 다국어 모델과의 성능 비교를 위해 12-레이어로 구성 후 [1]에서 제안한 학습 전략을 사용했다. 128 길이의 문장을 90만 step 학습 후 512 길이의 문장을 이용해 10만 step을 추가 학습하여 총 100만 step을 학습했다. Mecab 형태소 분석기를 사용한 모델이 WPM(WordPiece Model)을 적용한 BERT 다국어 모델보다 F1은 1.38%, EM은 12.4% 높았다.

표 2. 사전 구성 별 KorQuAD Dev 셋 성능 비교

모델	F1	EM	Step
BERT, multilingual cased	89.70%	70.21%	100만
BERT with Mecab	**91.08%**	**82.61%**	**100만**

4.2 Relation-aware Self-Attention 실험 및 결과

표 3, 4는 3.2절에서 제안한 Relation-aware Self-Attention의 실험 결과이다. 성능 비교를 위해 128길이 문장으로 구성 후 동일한 입력 포맷으로 각각 10만 step을 학습시켰다. 실험 결과, KorQuAD task에서 Relation-aware Self-Attention을 적용한 모델이 Self-Attention보다 F1 13.79%, EM 18.24% 높았으며, NSMC task에서는 정확도가 0.3% 높았다.

표 3. Self-Attention 별 KorQuAD Dev 셋 성능 비교

모델	F1	EM	Step
Self-Attention	72.83%	58.03%	10만
Relation-aware Self-Attention	**86.62%**	**76.27%**	**10만**

표 4. Self-Attention 별 NSMC Test 셋 성능 비교

모델	정확도	Step
Self-Attention	87.60%	10만
Relation-aware Self-Attention	**87.90%**	**10만**

그림 1은 Relation-aware Self-Attention과 Self-Attention의 학습 step 별 KorQuAD F1-스코어이다.

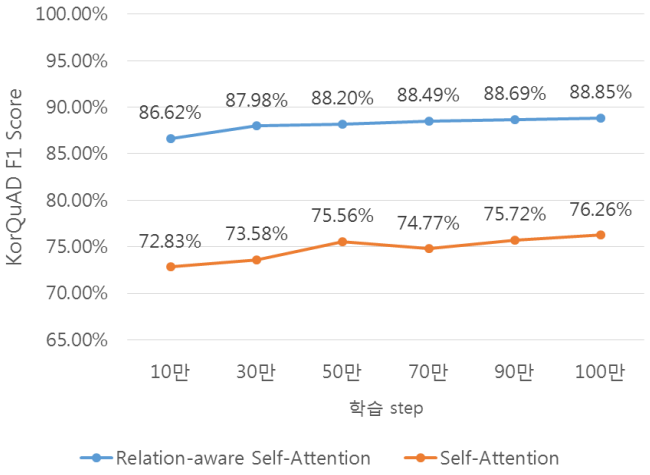

그림 1. 학습 Step 별 KorQuAD Dev 셋 성능

Relation-aware Self-Attention을 사용하여, 모델을 학습할 경우 낮은 학습 step에서도 높은 성능을 보이면서 빠르게 수렴했다.

4.3 결과

표 6은 본 논문에서 제안한 학습 방법을 적용한 BERT 모델을 한국어 KorQuAD 1.0 task에 적용한 실험 결과이다. [1]에서 공개한 BERT 다국어 모델보다 F1은 1.56% 낮았으나, EM은 8.52% 높은 성능을 보여주었다.

표 6. KorQuAD Dev 셋 모델 성능 결과

모델	F1	EM	step	레이어
BERT, multilingual cased	89.70%	70.21%	100만	12
한국어 BERT (our model)	**88.14%**	**78.73%**	**30만**	**3**

표 7은 NSMC에 적용한 실험 결과이다. BERT 다국어 모델보다 1.58% 높은 정확도 보여주었다.

표 7. NSMC Test 셋 모델 성능 결과

모델	정확도	step	레이어
BERT, multilingual cased	86.72%	100만	12
한국어 BERT (our model)	**88.30%**	**30만**	**3**

5. 결론

본 논문에서는 빠른 학습을 위한 한국어 BERT 모델 학습 방법을 제안하고 이를 KorQuAD 1.0, NSMC 1.0을 이용해 성능을 평가했다. 실험 결과, KorQuAD 1.0 task에서는 BERT 다국어 모델보다 F1은 1.56% 낮았으나, EM은

8.52% 높은 성능을 보여주었다. NSMMC 1.0에서는 1.58% 높은 정확도를 보여주었다. Relation-aware Self-Attention을 적용할 경우 낮은 학습 step과 적은 레이어에서도 BERT 다국어 모델과 비슷한 성능을 보여주었다.

본 논문에서는 Mecab 형태소 분석기를 통한 토크나이징 후 형태소 분석 결과를 이용해 사전을 구성했다. 그러나 사전 크기가 클수록 학습 속도가 늘어나는 문제가 있었다. 표 8은 사전 크기에 따른 학습시간을 보여준다. 향후에는 효과적으로 사전 크기를 줄일 수 있는 방법을 통해 모델 성능과 학습 속도를 향상하고자 한다.

표 8. 사전 크기에 따른 학습시간

모델	사전 크기	step	학습시간	TPU
Small BERT	32,000	10만	60분	16
	64,000	10만	80분	16
	128,000	10만	120분	16

참고문헌

[1] J. Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, 2018.

[2] Kuratov, Yuri, and Mikhail Arkhipov. "Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language." arXiv preprint arXiv:1905.07213 (2019).

[3] korBERT, http://aiopen.etri.re.kr

[4] Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).

[5] Sun, Yu, et al. "ERNIE: Enhanced Representation through Knowledge Integration." arXiv preprint arXiv:1904.09223 (2019).

[6] Liu, Yinhan, et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv:1907.11692 (2019).

[7] Joshi, Mandar, et al. "SpanBERT: Improving Pre-training by Representing and Predicting Spans." arXiv preprint arXiv:1907.10529 (2019).

[8] BERT with MECAB tokenizer for Korean text, https://github.com/yeontaek/BERT-MECAB-Korean-Model

[9] Mecab-ko, https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/

[10] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[11] Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. "Self-attention with relative position representations." arXiv preprint arXiv:1803.02155 (2018).

[12] 임승영, 김명지, and 이주열. "KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋." 한국정보과학회 학술발표논문집 (2018): 539-541.

[13] Naver sentiment movie corpus v1.0, https://github.com/e9t/nsmc