딥러닝 사전학습 언어모델 기술 동향

Recent R&D Trends for Pretrained Language Model

임준호 (J.H. Lim, joonho.lim@etri.re.kr) 언어지능연구실 책임연구원/PL
김현기 (H.K. Kim, hkk@etri.re.kr) 언어지능연구실 책임연구원
김영길 (Y.K. Kim, kimyk@etri.re.kr) 언어지능연구실 책임연구원/실장

ABSTRACT

Recently, a technique for applying a deep learning language model pretrained from a large corpus to fine-tuning for each application task has been widely used as a language processing technology. The pretrained language model shows higher performance and satisfactory generalization performance than existing methods. This paper introduces the major research trends related to deep learning pretrained language models in the field of language processing. We describe in detail the motivations, models, learning methods, and results of the BERT language model that had significant influence on subsequent studies. Subsequently, we introduce the results of language model studies after BERT, focusing on SpanBERT, RoBERTa, ALBERT, BART, and ELECTRA. Finally, we introduce the KorBERT pretrained language model, which shows satisfactory performance in Korean language. In addition, we introduce techniques on how to apply the pretrained language model to Korean (agglutinative) language, which consists of a combination of content and functional morphemes, unlike English (refractive) language whose endings change depending on the application.

KEYWORDS 딥러닝 사전학습 언어모델, BERT, RoBERTa, ALBERT, BART, ELECTRA, KorBERT

Ⅰ. 서론

딥러닝 사전학습 언어모델은 수십~수백 GB 이상의 대용량 텍스트 데이터로부터 언어의 문법 및 의미를 학습하여 다양한 응용 태스크에 범용적으로 적용하는 기술로, 기존 사전학습 언어모델을 사용하지 않는 방법 대비 우수한 성능을 보임이 많은 연구를 통하여 증명되었다.

2018년 10월 BERT(Bidirectional Encoder Representations from Transformers) 언어모델[1]이 공개된 이후, 대용량 텍스트 데이터로부터 더 효과적으로 언어모델을 학습하기 위한 여러 방법이 제안되었

* DOI: https://doi.org/10.22648/ETRI.2020.J.350302
* 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No. 2013-0-00131, (엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발].

으며, 본 고에서는 이와 같은 딥러닝 언어모델의 최근 기술 동향을 소개한다.

본 고의 구성은 다음과 같다. Ⅱ장에서 BERT 딥러닝 사전학습 언어모델의 동기, 학습방법, 모델에 대해 소개하고, Ⅲ장에서는 BERT 이후의 최신 언어모델 기술 동향에 대해 소개한다. Ⅳ장에서는 딥러닝 언어모델을 한국어에 적용한 KorBERT 언어모델에 대해 소개하고, Ⅴ장에서는 결론에 대해 소개한다.

Ⅱ. BERT 딥러닝 언어모델 기술 개요

1. 동기

언어는 심볼(Symbol)로 구성되나 딥러닝 뉴럴(Neural) 접근방법은 실수 값(Real value) 사이의 연산으로 표현되기 때문에, 딥러닝 기반 언어분석에 있어서 심볼을 실수 값으로 변환하는 워드 임베딩(Word embedding) 작업은 필수적이다.

기존 딥러닝 기반 언어처리 기술에서는 각 심볼을 미리 정의한 실수 벡터로 변환한 이후, 뉴럴 네트워크를 적용하는 접근 방법을 사용하였다[2,3]. 하지만 기존 방법의 경우 “하늘에서 눈이 내린다.”와 “졸려서 눈이 감긴다.”와 같이 동일한 심볼(“눈”)이 서로 다른 의미로 사용된 경우에도 모두 동일한 실수 벡터를 사용하게 된다는 문제가 있다.

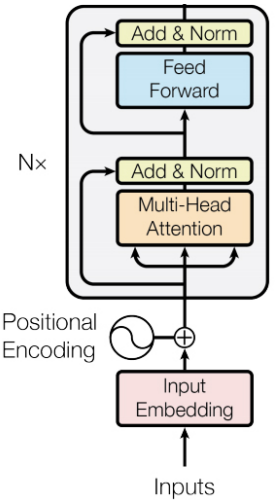
BERT 언어모델에서는 이와 같은 문제를 해결하기 위하여, 각 단어에 대해서 주변 단어들과의 자가집중(self-attention) 연산을 거친 결과 벡터를 해당 단어의 문맥을 표현한 벡터로 보고, 이를 응용 태스크에 적용하는 방법을 제안하였다.

2. 사전학습 태스크 및 모델 구조

BERT 언어모델의 입력과 출력은 N개의 단어(토큰)을 입력하여, 각 토큰(N개)의 벡터를 출력하는 구조를 가진다.

사전학습 태스크로는 Masked LM(MLM) 태스크와 Next Sentence Prediction(NSP) 태스크를 적용하였다. 첫 번째로, MLM 태스크는 주변 단어를 보고 해당 단어를 예측하는 태스크이다. 구체적으로, N개의 입력 문자열에 대하여 임의로 선택한 15%의 토큰을 [MASK] 토큰으로 변환한 후 BERT 언어모델에 입력하여, [MASK] 토큰으로 변환된 토큰의 출력 벡터를 이용하여 [MASK] 되기 이전의 단어를 예측하였다(실제적으로는 전체 [MASK] 토큰 중, 80%는 [MASK] 토큰을, 10%는 원본 단어를, 나머지 10%는 임의의 단어로 치환하는 방법을 적용하였다). 두 번째로, NSP 태스크는 두 문장 열 쌍이 선/후 관계가 맞는지를 학습한 태스크이다. 두 문장 열 쌍을 구분하기 위하여 N개의 입력을 “[CLS] segment_a [SEP] segment_b [SEP]”와 같이 구성하고, 두 segment 쌍이 연속 문장인지 여부를 “[CLS]” 토큰을 이용하여 이진 분류를 수행한다.

BERT 언어모델의 모델은 그림 1과 같이 다중 레



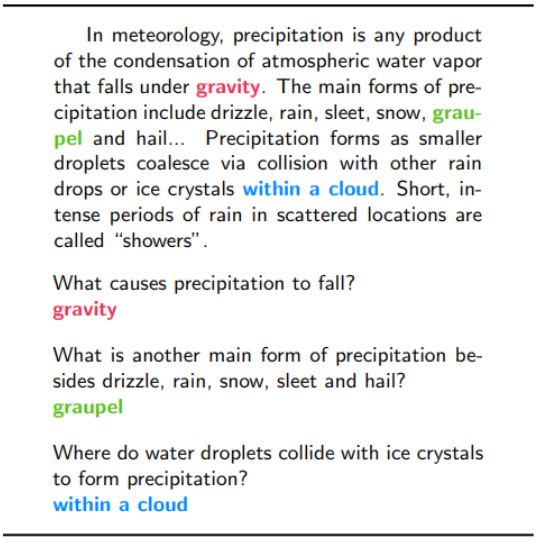
출처 Reprinted with author’s permission from https://arxiv.org/abs/1706.03762

그림 1 트랜스포머 인코더 구조

이어 트랜스포머(Transformer)의 인코더(Encoder) 구조를 사용한다[4].

BERT 모델의 입력이 되는 N개 토큰은 각 토큰 임베딩 외에 segment_a와 segment_b를 표현하는 segment 임베딩 및 각 토큰의 위치를 표현하는 position 임베딩과 결합되어 BERT 모델에 입력된다.

트랜스포머의 각 레이어의 동작은 다음과 같다. Multi-head Attention은 N개의 각 입력 토큰에 대해서 모든 가능한 토큰 조합(N×N) 사이의 가중치를 계산하고, 각 가중치에 따른 가중합 연산을 통하여 출력 벡터를 생성한다(Multi-head는 head 개수만큼 독립적인 출력 벡터를 생성한 후 concat 연산을 수행하는 것을 나타내고, 이때 concat된 출력 벡터의 크기가 입력 벡터의 크기와 같아지도록 개별 head의 출력 벡터 크기를 조정한다). 이후, 입력 벡터와의 원소합 및 레이어 정규화를 통하여 출력 층의 loss가 하위 계층까지 안정적으로 전파되도록 한다. 이후, 각 토큰에 대해서 2-layer FFNN 결과와 원소합을 통하여 벡터 표현을 확장한다.


In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

출처 Reprinted from https://www.aclweb.org/anthology/D16-1264/, CC BY 4.0.

그림 2 SQuAD 기계독해 예제

BERT 언어모델은 base 모델의 경우 12 layer, 12 multi-head, 768차원 토큰 벡터로 구성하고, BERT large 모델의 경우 24 layer, 16 multi-head, 1024차원 토큰 벡터로 구성하여 학습한다.

3. 학습 및 결과

BERT 언어모델은 학습을 위한 데이터로 BooksCorpus(800M 단어)와 영어 위키백과(2,500M 단어)를 사용하여 학습하였고, 크기는 약 16GB이다. 사전학습은 256 배치(batch)를 사용하여, 1M step 횟수만큼 학습하였다.

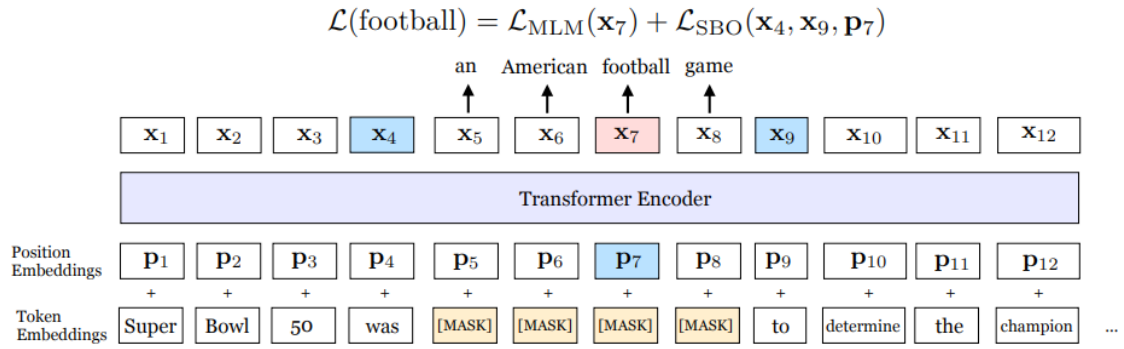
BERT 언어모델은 30,000개의 워드피스(Word-piece) 부분어휘 단위로 문장을 분할하여 학습하였다.

학습된 언어모델의 평가를 위하여 종합 언어이해 태스크(GLUE) 및 그림 2와 같은 기계독해 태스크(SQuAD v1.1, SQuAD v2.0) 등에 적용하여 평가하였다[5,6]. GLUE 평가 결과, 기존 최고 수준인 75.1점보다 높은 82.1점(BERTLarge, single model)을 기록하였고, SQuAD v2.0 평가 결과, 기존 최고 시스템 MIR-MRC F-Net(single model) 78.0%보다 5.1% 우수한 83.1%(BERTLarge, single model)의 F1 성능을 보였다.

III. 최신 딥러닝 언어모델 연구 동향

BERT 딥러닝 언어모델 이후, 대용량 텍스트 데이터로부터 언어모델을 더욱 잘 학습하기 위한 많은 연구가 제안되었으며, 세부적으로 ERNIE, Whole Word Masking, MASS, UniLM, XLNet, SpanBERT, RoBERTa, ALBERT, BART, ELECTRA, UniLMv2 등이 있다[7-16].

본 절에서는 주요 및 최신 연구를 중심으로,



출처 Reprinted with from M. Joshi, "SpanBERT: Improving Pre-training by Representing and Predicting Spans," Accepted at TACL, 2020.

그림 3 SpanBERT의 Span Boundary Objective 예제

SpanBERT, RoBERTa, ALBERT, BART, ELECTRA 기술에 대해 소개한다.

1. SpanBERT

SpanBERT는 2019년 7월 워싱턴대, 프린스턴대, AllenAI 연구소, 페이스북에서 수행한 연구이다 [11]. SpanBERT의 주요 개선 내용은 Span Masking, Span Boundary Objective, Single-sequence training의 세 가지이다.

첫 번째, Span masking은 BERT에서 각 토큰에 대해서 개별적으로 [MASK] 토큰을 변환한 것과 달리, 연속된 span 단위로 각 토큰에 대해서 [MASK] 토큰을 변환하였다. 연속된 span은 각 토큰에 대해 베르누이 트라이얼(Bernoulli trials)을 연속적으로 수행하였을 경우의 geometric distribution(p=0.2)을 사용하여 마스킹을 수행하였으며, 총 마스킹 토큰은 전체 토큰 수의 15%를 동일하게 사용하였다.

두 번째, Span Boundary Objective는 [MASK]로 변환된 단어의 원 단어 토큰을 추론할 때, 해당 단어의 출력 벡터로부터 추론하는 loss 외에 span의 경계에 위치한 단어의 출력 벡터로부터 해당 단어를 추론하는 loss를 추가하여 학습한 것이다. 그림 3을 살펴보면, 7번째 단어가 football일 때, x7 출력 토큰을 이용하여 football을 예측하는 loss와 span의 경계에 해당하는 x4 및 x9와 해당 단어의 위치인 p7을 이용하여 football을 예측하는 loss를 결합하여 사용한다.

세 번째, BERT 언어모델이 두 segment를 결합하여 NSP 태스크를 수행한 것과 달리, 단일 segment로 입력을 구성하고, NSP 태스크를 제외하여 학습하였을 경우, 성능이 개선됨을 보였다.

SpanBERT 언어모델은 BERTLarge 모델과 동일한 모델을 사용하였으며, 학습데이터는 BERT와 동일한 BookCorpus 및 영어 위키백과를 사용하여 학습하였다. 사전학습은 256 batch를 사용하여, 2.4M step 횟수만큼 학습하였다.

기계독해 데이터셋 대상 실험 결과, SQuAD v1.1 평가셋에서 Google BERT 91.3%보다 3.3%p 우수한 94.6%의 성능을 보였으며, SQuAD v2.0 평가셋에서 Google BERT 83.3%보다 5.4%p 우수한 88.7%의 성능을 보였다.

2. RoBERTa

RoBERTa는 2019년 7월 워싱턴대 및 페이스북에

서 공개한 논문으로, BERT 모델 구조를 더 견고(Robust)하게 학습할 수 있는 방법을 제안하였다[12]. RoBERTa에서 개선한 내용은 dynamic masking, input format, large batch 학습으로 구분할 수 있다.

첫 번째, dynamic masking은 매 학습 단계마다 15% 임의 마스킹 변환을 새로 적용한 방법으로, BERT 모델의 경우 미리 정의한 횟수(10회) 만큼의 랜덤 학습 데이터를 생성한 이후, 해당 데이터를 고정하여 학습에 반복적으로 사용하였음을 이야기하였으며, 더 많은 학습 데이터로 더 많은 step을 학습하기 위해서는 dynamic masking이 중요함을 이야기하였다.

두 번째, input format에서는 BERT와 같은 segment_pair + NSP 방법, 두 sentence에 대한 sentence_pair + NSP 방법, 단일 문서 내 연속 sentence 구성 (NO NSP) 방법, 문서 경계에 관계없이 연속 sentence 구성 (NO NSP) 방법을 비교 실험하였다. 실험 결과, sentence 단위로 사전학습을 수행할 경우 성능에 안 좋은 영향을 미치며, segment_pair + NSP 방법보다 연속된 문장 열 기반으로 NSP 태스크를 적용하지 않는 것이 성능 개선에 도움이 됨을 보였다.

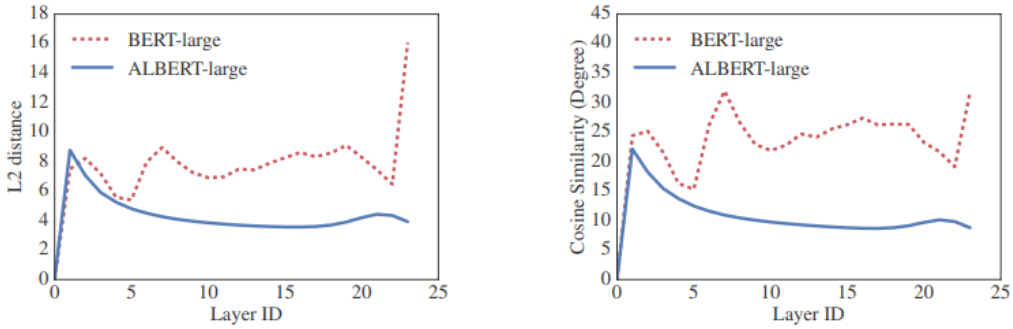
마지막으로, 사전학습 언어모델에 있어서 기존 256 배치 크기보다 큰 2K 배치 또는 8K 배치를 사용하는 것이 성능 개선에 도움이 됨을 보였다.

RoBERTa 모델의 실험 결과는 참고문헌 [12] 논문의 표 2와 같다. BERT 모델과 같이 BooksCorpus와 영어 위키백과를 사용하여 8K 배치로 100K step 학습하였을 경우, SQuAD 2.0 F1 기준 87.3%이나, 학습데이터를 CC-NEWS, OPENWEBTEXT, STORIES 데이터를 포함하여 160G로 늘리고 500K까지 학습하였을 경우, SQuAD v2.0 기준 F1이 89.4%까지 향상되었으며, 이는 BERT 및 XLNet 등 다른 기존 연구보다 우수한 성능임을 보였다.

3. ALBERT

ALBERT는 2019년 9월 Google에서 공개한 논문으로, BERTLarge보다 큰 모델을 효과적으로 학습하여 성능을 개선하기 위한 방법을 연구한 논문이다[13].

초기 실험으로 24 layer의 BERTLarge 모델을 1024 차원 벡터에서 2048차원 벡터로 크기를 키웠을 경우, RACE 기계독해 평가셋에서 BERTLarge의 모델은 73.9%의 성능을 보이나, 2048차원으로 학습한 모델은 54.3%로 성능이 크게 하락함을 보이고, 더 큰 모델을 효과적으로 학습하기 위한 방법을



출처 Reprinted with author's permission from https://arxiv.org/abs/1909.11942

그림 4 ALBERT 레이어 parameter sharing 효과

제안하였다. ALBERT에서 제안한 방법은 factorized embedding parameterization, cross-layer parameter sharing, inter-sentence coherence loss 의 세 가지이다.

첫 번째, factorized embedding parameterization은 BERT 모델의 동기와 같이 토큰 임베딩 파라미터는 주변 문맥에 독립적인 파라미터이고, 트랜스포머 레이어 내의 히든 벡터는 주변 문맥을 반영한 파라미터이다. 따라서 주변 문맥에 독립적인 파라미터인 토큰 임베딩 파라미터의 차원을 128차원으로 축소하고, 128차원의 벡터를 FFNN을 이용하여 트랜스포머 레이어에서 사용하는 차원으로 변환하여 모델 입력으로 사용하였다. 이를 통하여 토큰 임베딩 파라미터 수를 1024×vocab_num에서 (128×vocab_num + 128×1024)로 크게 줄일 수 있다.

두 번째, cross-layer parameter sharing은 트랜스포머의 각 레이어에 포함된 학습 파라미터(multi-head attention 및 FFNN 파라미터)를 모든 레이어에 동일하게 적용한 방법이다. 그림 4와 같이 각 레이어의 입력 벡터와 출력 벡터 사이의 L2 distance 및 cosine 유사도를 계산한 결과, parameter sharing을 사용한 경우가 그렇지 않은 경우보다 안정적(Smooth)인 변화를 보임을 확인하였다.

세 번째, inter-sentence coherence loss는 기존 NSP 태스크가 segment의 주제(topic)이 다름을 인식하는 문제와 segment가 일관됨(Coherence)을 인식하는 문제가 혼합되어 있음을 지적하고, 주제 인식 문제를 제거하기 위한 Sentence-order prediction(SOP) 태스크를 제안하였다. SOP 태스크는 동일 문서에서 연속적으로 추출된 segment_a와 segment_b에 대해서 50%는 원래 순서로 입력하고, 50%는 순서를 바꿔서 입력하여, 순서가 바뀌었는지 여부를 인식하는 태스크이다.

ALBERT 언어모델을 RoBERTa와 같이 160G 데이터로 학습하고, 4K 배치 크기에, 1.5M step 학습하였다. SQuAD v2.0 테스트셋 대상 평가 결과, Single model은 RoBERTa 모델 89.8% F1보다 1.1% 우수한 90.9% F1 성능을 보였으며, ensemble 모델은 92.2% F1 성능을 보였다.

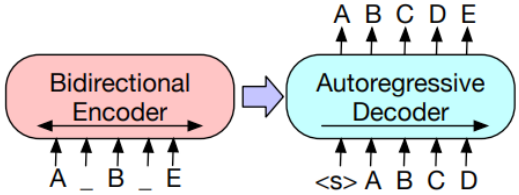
4. BART

BART(Bidirectional and Auto-Regressive Transformers)는 2019년 10월 페이스북에서 공개한 논문이다[14]. 기존 BERT 및 후속 연구가 트랜스포머의 인코더(Encoder)만을 사용한 것과 달리, 인코더와 디코더(Decoder)를 같이 사용하여 언어모델을 학습하였고, 디코더에서는 이전 단계의 출력 결과를 다음 단계의 입력으로 사용(Auto-regressive)하였다.

트랜스포머 인코더만 사용할 경우, 입력 토큰 수와 출력 토큰 수가 같다는 제약이 따르지만, BART의 경우 인코더와 디코더를 같이 사용하여 입력 토큰 수와 출력 토큰 수가 다른 경우에도 학습이 가능하다는 장점이 있다.

BART는 사전학습 태스크로 token masking, token deletion, text infilling, sentence permutation, document rotation의 사전 학습 태스크를 제안하였으며, 이 중 text infilling은 임의 길이의 텍스트 span을 하나의 [MASK] 토큰으로 변환하는 것이다. 변환 대상 텍스트 span은 단위 시간당 사건의 발생 확률을 모델링한 포아송 분포(Poisson distribution)를 사용하였으며, λ=3을 사용하였다. Text infilling 시, 0-길이 span도 [MASK]로 변환하여 해당 위치에 변환된 토큰이 없음도 인식하도록 모델을 학습하였다. 그림 5는 BART의 text infilling 수행 예를 보여준다. A 다음의 _는 0 길이 text infilling 예제이고, B 다음의 _에서 C와 D 토큰을 복원하는 예제이다.

BART 논문에서는 제안한 트랜스포머 인코더-



출처 Reprinted with author's permission from https://arxiv.org/abs/1910.13461

그림 5 BART 모델 text infilling 예제

디코더 구조 외에, MASS, GPT, XLNet, UniLM 모델 구조와 비교 실험을 통하여 제안한 방법의 우수함을 보였다.

BART 모델은 BERT의 24 layer를 인코더 12 layer, 디코더 12 layer로 분할하여 모델을 구성하였고, RoBERTa와 같은 160GB의 학습데이터를 사용하였다. 학습에 사용한 배치 크기는 8000이고, 학습 step은 500K를 사용하였다.

실험 결과, BART는 언어이해 태스크에서 RoBERTa와 비슷한 수준의 성능을 보였으며, 기존 BERT 및 RoBERTa에서 적용이 어려운 언어 생성(Generation) 태스크에서도 우수한 성능을 보였다. 언어이해 태스크인 기계독해 SQuAD v1.0에서는 RoBERTa와 동일한 F1 성능을 보였으며, SQuAD v2.0에서는 RoBERTa와 0.2% 차이인 89.2% F1 성능을 보였다. 언어 생성 태스크인 XSum 요약 태스크에서는 선행 연구 최고 성능인 31.27 Rouge-L 점수보다 5.98점 높은 37.25 Rouge-L 점수를 획득하였다.

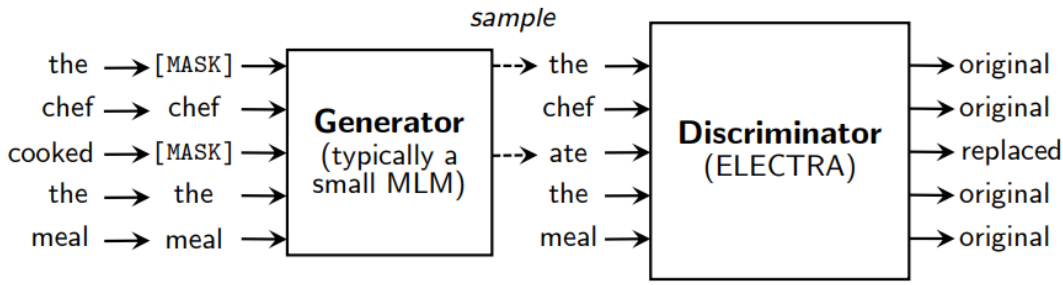
5. ELECTRA

ELECTRA는 스탠포드대 및 Google Brain에서 수행한 연구로, ICLR 오픈리뷰 사이트에 공개하여 ICLR 2020에 출판 예정인 연구이다[15].

ELCTRA에서 해결하고자 한 문제는 N개의 입력 토큰 중 [MASK]로 변환된 15%에서만 loss 값이 계산되기 때문에 언어모델의 학습 효율성이 떨어진다는 점이다. 이를 해결하기 위하여 N개의 입력 토큰 전체에서 loss를 계산하기 위한 방법을 제안하였다.

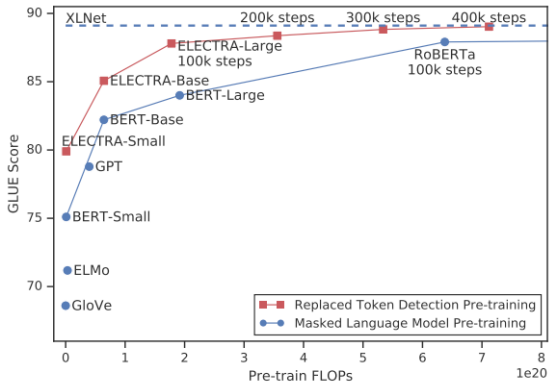
ELECTRA 모델 학습 예제는 그림 6과 같다. 우선, BERT와 동일하게 15%의 토큰에 대해서 [MASK]로 변환한다. 이후, Generator 뉴럴넷을 이용하여 각 [MASK] 토큰에 대해서 예측 단어를 추론하여 다시 변환한다. 마지막으로, BERT와 동일한 Discriminator에서 Generator에서 변환한 입력을 받아 모든 토큰에 대해 original 또는 replaced 여부를 분류한다.

ELECTRA는 Generator 및 Discriminator 구조를 사



출처 Reprinted with author's permission from https://openreview.net/forum?id=r1xMH1BtvB

그림 6 ELECTRA 모델 학습 예제



출처 Reprinted with author's permission from https://openreview.net/forum?id=r1xMH1BtvB

그림 7 ELECTRA 모델 학습 속도

용한다는 점에서 GAN(Generative Adversarial Networks)과 유사한 구조를 가지지만, Generator에서 원래 단어와 동일한 단어를 생성할 경우 original로 분류하고, 목적 함수가 Discriminator를 속이기 위한 loss가 아닌 Maximum Likelihood를 위한 Loss라는 점에서 GAN과 차이를 가진다.

ELCTRA 모델은 처리 텍스트의 15%가 아닌 전체 텍스트로부터 loss가 계산되기 때문에 초기학습 속도가 빠르다는 장점을 가진다. 그림 7과 같이 동일한 사전학습 FLOPs를 처리하였을 경우, BERT-Base, BERT-Large, RoBERAa보다 우수한 성능을 보인다.

ELECTRA의 학습데이터는 XLNet과 동일한 데이터를 사용하였고 약 142GB이다. 배치 크기는 2048이고, 1.75M step 학습을 수행하였으며, 이는 BERT_Large 모델 대비 약 4.4배 학습을 수행한 결과이며, RoBERTa(4.5배)와 비슷한 수준의 학습 결과이다.

실험 결과는 그림 8과 같다. SQuAD 2.0 테스트 셋 기준, RoBERTa 모델 89.8% F1, ALBERT 90.9% F1보다 우수한 91.4% F1 성능을 보였다. 참고로 ALBERT는 ELECTRA 모델보다 10배 많은 학습 연산을 수행하였다.

IV. 한국어 딥러닝 언어모델 KorBERT

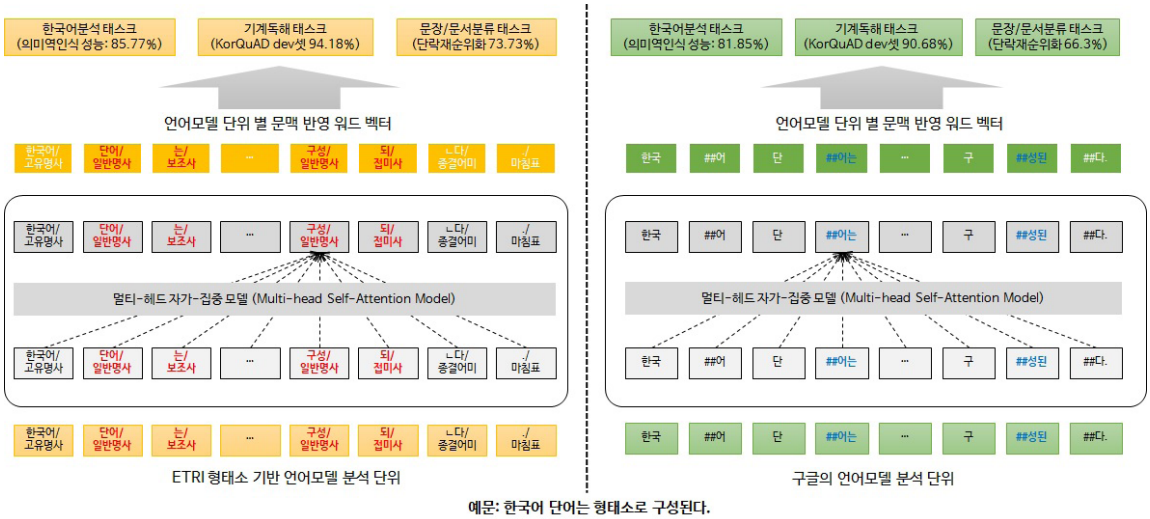
본 절에는 ETRI에서 개발한 한국어 딥러닝 언어모델 KorBERT에 대해 소개한다[17].

딥러닝 언어모델은 문장을 토큰 단위로 구분한 후, 토큰과 토큰 사이의 관계를 학습하는 모델이기

Model	Train FLOPs	Params	SQuAD 1.1 dev		SQuAD 2.0 dev		SQuAD 2.0 test	
			EM	F1	EM	F1	EM	F1
BERT-Base	6.4e19 (0.09x)	110M	80.8	88.5	–	–	–	–
BERT	1.9e20 (0.27x)	335M	84.1	90.9	79.0	81.8	80.0	83.0
SpanBERT	7.1e20 (1x)	335M	88.8	94.6	85.7	88.7	85.7	88.7
XLNet-Base	6.6e19 (0.09x)	117M	81.3	–	78.5	–	–	–
XLNet	3.9e21 (5.4x)	360M	**89.7**	**95.1**	87.9	**90.6**	87.9	90.7
RoBERTa-100K	6.4e20 (0.90x)	356M	–	94.0	–	87.7	–	–
RoBERTa-500K	3.2e21 (4.5x)	356M	88.9	94.6	86.5	89.4	86.8	89.8
ALBERT	3.1e22 (44x)	235M	89.3	94.8	87.4	90.2	88.1	90.9
BERT (ours)	7.1e20 (1x)	335M	88.0	93.7	84.7	87.5	–	–
ELECTRA-Base	6.4e19 (0.09x)	110M	84.5	90.8	80.5	83.3	–	–
ELECTRA-400K	7.1e20 (1x)	335M	88.7	94.2	86.9	89.6	–	–
ELECTRA-1.75M	3.1e21 (4.4x)	335M	**89.7**	94.9	**88.0**	**90.6**	**88.7**	**91.4**

출처 Reprinted with author's permission from https://openreview.net/forum?id=r1xMH1BtvB

그림 8 ELECTRA 모델 실험 결과



출처 http://aiopen.etri.re.kr

그림 9 KorBERT 언어모델 단위 예제

때문에 올바른 토큰의 구분이 중요하다.

영어는 굴절어로 활용에 따라 어미가 변화하는 언어이지만, 한국어는 교착어로 명사/동사와 같은 내용어와 조사/어미와 같은 기능어가 결합하여 하나의 어절을 구성하는 언어이기 때문에, 올바른 문맥표현을 위해서는 내용어와 기능어를 구분하는 단계가 필요하다.

KorBERT 언어모델에서는 입력 문장에 대해 형태소 분석을 수행하고, 형태소 분석된 결과에 기반하여 각 토큰 간의 문맥표현을 학습한다. 그림 9의 오른쪽과 같이 형태소 분석을 수행하지 않았을 경우 "단어는" 어절이 "단+어는"으로, "구성된다" 어절이 "구+성된+다"로 구분되는 것을 확인할 수 있으나, 그림 9의 왼쪽과 같이 형태소 분석을 수행하였을 경우, 각 어절이 내용어와 기능어로 올바르게 구분됨을 확인할 수 있다.

KorBERT 언어모델 학습을 위하여 백과사전류 텍스트와 약 15년 분량의 신문기사를 수집하여 23GB의 대용량 텍스트를 대상으로, 47억 개의 형태소를 학습하였다.

딥러닝 언어모델 학습방법으로 BERT 사전학습 언어모델의 MLM 방법 대비 사전학습 태스크 및 가중치 학습방법을 개선하여 적용하였다.

KorBERT 언어모델은 형태소 분석 적용 여부에 따른 응용 태스크별 성능 개선 정도를 측정하기 위하여, 형태소 분석 결과를 기반으로 사용하는 KorBERT-Morphology 모델과 형태소 분석을 수행하지 않고 End-to-End로 적용 가능한 KorBERT-Wordpiece 모델을 학습 및 공개하였다.

KorBERT 언어모델 평가는 다음 5개를 대상으로 수행하였다.

- 의미역 인식(Semantic Role Labeling): 문장 내에서 술어에 의해 기술되는 사건에 대한 개체들의 역할을 인식
- 기계 독해(Machine Reading Comprehension): 주어진 단락에서 질문이 요구하는 정답을 찾음
- 단락 순위화(Passage Ranking): 검색결과 집합에서 질문에 찾는 정답이 들어 있는 단락 순위화
- 문장 유사도 추론(Natural Language Inference):

구분	의미역인식	기계독해	단락순위화	문장유사도추론	문서주제분류
평가데이터 및 규격	Korean Propbank, 학습: 19,302 문장 평가: 3,773 문장	KorQuAD 데이터, 학습: 60,406건 평가: 5,773건 (dev셋)	학습: 45,521 질문 평가: 1,000 질문 (질문당 평균 8.7개 단락)	학습: 10,874문장쌍 평가: 1,209문장쌍 (이진 분류체계: 유사, 무관)	학습: 9,301건 평가: 1,035건 (54개 분류체계)
평가 방법	F1[2]	Exact Match[3] / F1	Precision@Top1	Accuracy	Accuracy
(Google) Word Piece[4]기반 한국어 언어모델	81.85%	80.82% / 90.68% (정답 경계 구분을 위해 후처리 수행)	66.3%	79.4%	91.1%
(엑소브레인) Word Piece 기반 한국어 언어모델	85.10%	80.70% / 91.94% (정답 경계 구분을 위해 후처리 수행)	70.5%	82.7%	93.4%
(엑소브레인) 형태소 기반 한국어 언어모델	85.77%	86.40% / 94.18%	73.7%	83.4%	93.7%

출처 http://aiopen.etri.re.kr

그림 10 5개 태스크 대상 KorBERT 언어모델 실험 결과

2개 문장 간 의미가 동일한지 여부를 분류

- 문서 주제분류: 문서의 주제를 기 정의된 54개의 클래스 중 하나로 분류

그림 10의 실험결과와 같이, KorBERT 언어모델은 구글에서 배포한 다국어 모델 대비 우수한 성능을 보임을 확인할 수 있으며, 한국어의 경우 형태소 분석을 수행한 KorBERT-Morphology 모델이 KorBERT-Wordpiece 모델보다 모든 태스크에서 우수함을 확인하였다.

V. 결론

본 고에서는 딥러닝 언어처리를 위한 사전학습 언어모델의 기술 동향에 대해 살펴보았다. BERT 언어모델이 제안된 이후, 언어모델 학습을 개선하기 위한 많은 연구가 제안되었으며, 최근에는 트랜스포머의 인코더-디코더 구조를 활용한 연구 및 GAN과 유사한 Generator-Discriminator 구조를 활용한 언어모델 학습 연구가 제안되었다.

언어처리 분야의 향후 연구로 현재 트랜스포머 인코더 모델의 효율성 개선, 대용량 텍스트로부터 학습하는 지식의 확장, 외부 지식/메모리 활용, 인식 결과의 설명 가능성 등의 연구가 수행될 것으로 예상되며, 향후 연구에서도 대용량 텍스트로부터 지식을 사전에 학습하는 사전학습 언어모델 접근방법은 계속하여 유효한 접근방법일 것으로 예상된다.

용어해설

딥러닝 사전학습 언어모델 대용량 텍스트로부터 각 토큰 사이의 문법/의미적 관계를 학습한 딥러닝 언어모델로, 응용 태스크에 적용 시 다양한 태스크에서 우수한 성능을 보임

트랜스포머 2017년 6월 기계번역을 위해 제안된 모델로, 자가집중(self-attention) 방법을 이용하여 원문 인코딩 및 번역문 디코딩을 수행한 모델

약어 정리

BERT	Bidirectional Encoder Representations from Transformers
GLEU	General Language Understanding

	Evaluation benchmark
KorBERT	Korean BERT Model
SQuAD	Stanford Question Answering Dataset

참고문헌

[1] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. North Am. Association Computat. Linguistics (NAACL)-HLT, Minneapolis, MN, USA, June 2-7, 2019, pp. 4171-4186.

[2] T. Mikolov et al., "Distributed representations of words and phrases and their compositionality," in Proc. Int. Conf. Neural Inf. Process. Syst., 2013, pp. 3111-3119, doi: 10.5555/2999792.2999959.

[3] P. Bojanowski et al., "Enriching word vectors with subword information," Trans. Assoc. Comput. Linguistics, vol. 5, Dec. 2017, pp. 135-146.

[4] A. Vaswani et al., "Attention is all you need," in Proc. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017, pp. 30-34.

[5] https://gluebenchmark.com/

[6] https://rajpurkar.github.io/SQuAD-explorer/

[7] Y. Sun et al., "ERNIE: Enhanced Representation through Knowledge Integration," arXiv preprint arXiv:1904.09223, 2019.

[8] K. Song et al., "Mass: Masked sequence to sequence pre-training for language generation," in Int. Conf. Mach. Learning, Long Beach, CA, USA, 2019, pp. 5926-5936.

[9] L. Dong et al., "Unified language model pre-training for natural language understanding and generation," arXiv preprint arXiv:1905.03197, 2019.

[10] Z. Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," arXiv preprint 1906.08237, 2019.

[11] M. Joshi et al., "SpanBERT: Improving pre-training by representing and predicting spans," arXiv preprint 1907.10529, 2019.

[12] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

[13] Z. Lan et al., "ALBERT: A Lite BERT for Selfsupervised Learning of Language Representations," in Int. Conf. Learning Representations, Addis Ababa, Ethiopia, May 2020.

[14] M. Lewis et al., "Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461, 2019.

[15] K. Clark et al., "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." in Int. Conf. Learning Representations, Addis Ababa, Ethiopia, May 2020.

[16] H. Bao et al., "UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training." arXiv preprint arXiv:2002.12804, 2020.

[17] http://aiopen.etri.re.kr/service_dataset.php