『 [#BERT](https://www.facebook.com/hashtag/bert?source=note&epa=HASHTAG) [#KorQuAD](https://www.facebook.com/hashtag/korquad?source=note&epa=HASHTAG)

안녕하세요. 첨글을 올리네요. ETRI 한국어 BERT모델에 공개형 mecab 형태소분석기를 적용한 결과를 공개합니다.

얼마 전에 ETRI에서 한국어 BERT 모델을 공개했고, 솔트룩스에서 개발한 형분석기로 진행한 것도 꽤 괜찮은 결과가 나온것으로 포스팅되었습니다.(<https://www.facebook.com/groups/TensorFlowKR/permalink/926295594378166/> )

그래서, 저는 일반적으로 많이 사용하는 공개 형분석기인 mecab으로는 어떤 정도수준으로 나올까 해서 ETRI에서 공개한 BERT모델에 mecab형분석기를 이용하여 테스트를 했습니다.

mecab형분석기를 적용하는 방법은, vocab.korean\_morp.list인 vocab파일의 태킹부분을 수정해야 합니다. mecab은 태킹이 세종품사표시방식과 약간 틀린 부분이 있어서 (ex: 붙임표가 세종품사는 SO인데 mecab은 SY) 직접 수정을 하야 합니다. 하지만 수정할 게 실제 그렇게 많지 않습니다. 10분내외로 완료가능합니다.

그리고, infect가 되어있는 부분을 나누어 주는 부분을 run\_mrc\_finetunning\_sample.py에 수정 추가하여 주면 됩니다.

ex) mecab분석기의 inflect한 형태 (한/XSA+ETM -> 한/XSA\*+ㄴ/ETM/\*)

한가한

한가 XR,\*,F,한가,\*,\*,\*,\*

한 XSA+ETM,\*,T,한,Inflect,XSA,ETM,하/XSA/\*+ㄴ/ETM/\*

EOS

※ 소스공개하면 쉽긴한데 ETRI에서 모델 다운받을때 서약서에 ETRI 허락을 받아야 한다고 해서 그냥 말로 설명합니다.

korquad보드에는 올리는 것은 복잡고 귀찮아서... 그냥 dev set 기준으로 평가해서 (em: 84.8, f1: 92.99)라는 결과를 얻었습니다.

솔트룩스 형분석기가 7위정도니 약 10위정도 할 수 있겠네요. 』

※ 한국어 BERT 모델 (BERT : Bidirectional Encoder Representations from Transformers)

: 자연언어 처리 태스크를 교육 없이 양방향으로 사전학습하는 첫 시스템이다. 교육 없음이란 BERT가 보통의 텍스트 코퍼스만을 이용해 훈련되고 있다는 것을 의미한다. 이것은 웹 상에서 막대한 양의 보통 텍스트 데이터가 여러 언어로 이용 가능하기 때문에 중요한 특징으로 꼽는다.

일부 성능 평가에서 인간보다 더 높은 정확도를 보이며, 자연 언어 처리(NLP) AI의 최첨단 딥러닝 모델이다.

언어표현 사전학습의 새로운 방법으로 그 의미는 ‘큰 텍스트 코퍼스(ex) Wikipedia)를 이용하여 범용목적의 ‘언어 이해(language understanding)’ 모델을 훈련시키는 것과 그 모델에 관심 있는 실제의 자연 언어 처리 태스크(질문, 응답 등)에 적용하는 것이다.

BERT는 문맥에 의존하는 특징적인 표현의 전학습을 실시하는 대응을 바탕으로 구축되었다.

## 참고 : <http://docs.likejazz.com/bert/>

※ mecab 형태소 분석기

: 일본어 형태소 분석기 mecab을 의미한다.

Mecab-ko는 이 mecab을 한국어로 사용 가능하도록 포팅한 버전으로 C++로 개발된 것이다. 그리고 속도 및 성능이 전반적으로 좋다.

※ 한가한

한 = 하(XSA = 형용사 파생 접미사) + ㄴ(ETM = 관형형 전성 어미)

Mecab-ko에서는 위 처럼 형태소 분석을 실시 한다.

『 [#BERT](https://www.facebook.com/hashtag/bert?source=feed_text&epa=HASHTAG) [#KorQuAD](https://www.facebook.com/hashtag/korquad?source=feed_text&epa=HASHTAG)

안녕하세요 TF Korea!!

얼마 전에 ETRI에서 한국어 BERT 모델을 공개했는데요,

저희 팀에서 간단하게 KorQuAD 테스트를 해봐서 결과를 공유드립니다 https://static.xx.fbcdn.net/images/emoji.php/v9/t4c/1/16/1f642.png:-)

ETRI BERT는 원형 복원 된 형태소 단위로 tokenizing이 이루어 지는데, ETRI 형분석기는 일 5,000건으로 제한이 있어서

저희 솔트룩스에서 개발한 형분석기를 이용해 테스트를 수행하였습니다. https://static.xx.fbcdn.net/images/emoji.php/v9/t4c/1/16/1f642.png:-)

Dev set의 경우엔 ETRI에서 공개한 성능과 유사하게 나왔었고 (em: 85.10, f1: 93.34), test set으로는 7위에 랭크됐네요.. https://static.xx.fbcdn.net/images/emoji.php/v9/t4c/1/16/1f642.png:-)

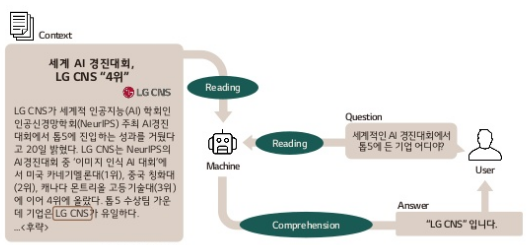
BERT는 사이즈가 너무 커서 학습할 엄두가 안나는 모델 중 하나인데,

감사하게도 ETRI에서 한국어 BERT를 공개해주셔서 정말 간단하게, 누구든지 테스트가 가능하게 됐네요!

앞으로도 자연어 분야에서 활발한 지식 공유가 이루어지면 좋겠습니다 ^^ 』

※ KorQuAD

: 한국어 Machine Reading Comprehension을 위해 만든 dataset입니다. 즉, 한국어 질의응답 과제를 위한 대규모 dataset이다.



Machine Reading Comprehension : 기계가 주어진 문단에 대해서 질문을 읽고 답변을 찾아내는기법

## 참고 : <https://www.youtube.com/watch?v=ntGwv6Ifoe8&feature=youtu.be&fbclid=IwAR1lXoz-wj9yAlkJhRKh-MN40hcT1sOSSIcHuGUzq5bE7CbCFha7TUsqgW8>

※ Tokenizing(토큰 생성)

: 자연어 문서를 분석하기 위해서 우선 긴 문자열을 분석을 위한 작은 단위로 나누어야 한다. 이 문자열 단위를 토큰(token)이라고 하고 이렇게 문자열을 토큰으로 나누는 작업을 Tokenizing이라고 한다. 영문의 경우에는 문장, 단어 등을 토큰으로 사용하거나 정규 표현식을 쓸 수 있다