RoBERTa를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱

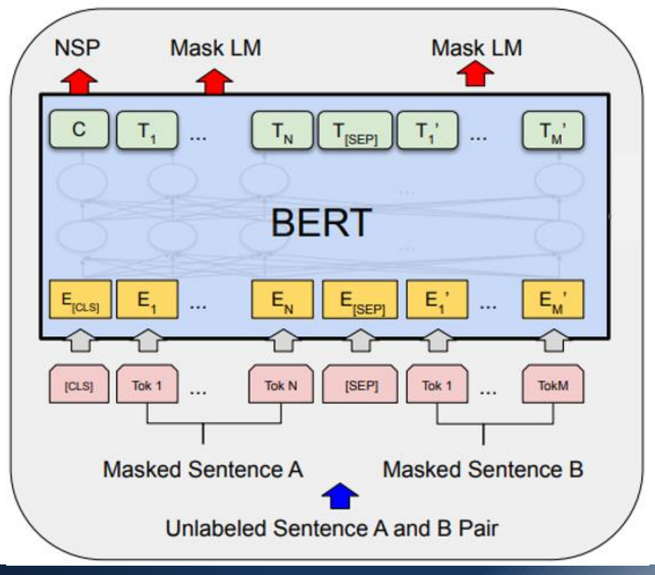
Cognitive Computing Lab.

민진우[O1], 나승훈[2], 신종훈[3], 김영길[4]
[12] 전북대학교, [34] 한국전자통신연구원
jinwoomin4488@gmail.com, nash@jbnu.ac.kr, {jhshin82, kimyk}@etri.re.kr

Ⅰ. 서론

BERT 양방향 트랜스포머 기반의 언어 모델로 양방향의 문맥을 보고 문장 내의 임의의 단어를 마스킹 (Masking)하고 예측하는 Masked LM 태스크와 추가로 A문장과 B 문장을 입력으로 받아 B문장이 A문장의 다음으로 적절한 문장인지 판별하는 NSP(Next Sentence Prediction) 태스크에 대한 언어 모델을 학습.

• 다양한 자연어 처리 테스크에서 성능 향상을 가져옴



Ⅱ. 한국어 RoBERTa 모델

Proposed System

본 논문에서는 형태소 기반 토크나이저와 BPE(Byte-Pair-Encoding) 기반의 토크나이저를 결합하여 미등록어에 강건한 하이브리드 토크나이저를 제안하고 한국어 RoBERTa 모델을 학습한 후 다양한 자연어 처리 태스크에 적용하여 기존의 BERT의 성능을 더욱 향상

ⅰ 하이브리드 토크나이저

- 형태소 토큰 5만개와 BPE 토큰 2만개를 단어장으로 구성
- 분석된 형태소를 형태소 단위를 우선적으로 단어장에서 매칭한 후 해당 형태소가 미등록어일 경우 형태소를 BPE 단위로 토크나이징 하는 하이브리드 방식을 사용

원문
고전주의와 바로크는 공통의 면이 있다.
형태소 분석 결과
고전주의/NNG, 와/JC, 바로크/NNG, 는/JX, 공통/NNG, 의/JKG, 면/NNG, 이/JKS, 있/VV, 다/EF, ./SF
토크나이징 결과
_고전, 주의, 와/JC, 바로크/NNG, 는/JX, 공통/NNG, 의/JKG, 면/NNG, 이/JKS, 있/VV, 다/EF, ./SF

단어장에 존재하지 않는 미등록 형태소 “고전주의/NNG”에 대해서는 품사를 잘라낸 형태소 “고전주의”가 BPE 모델을 통해 “_고전”, “주의”로 토큰화됨.

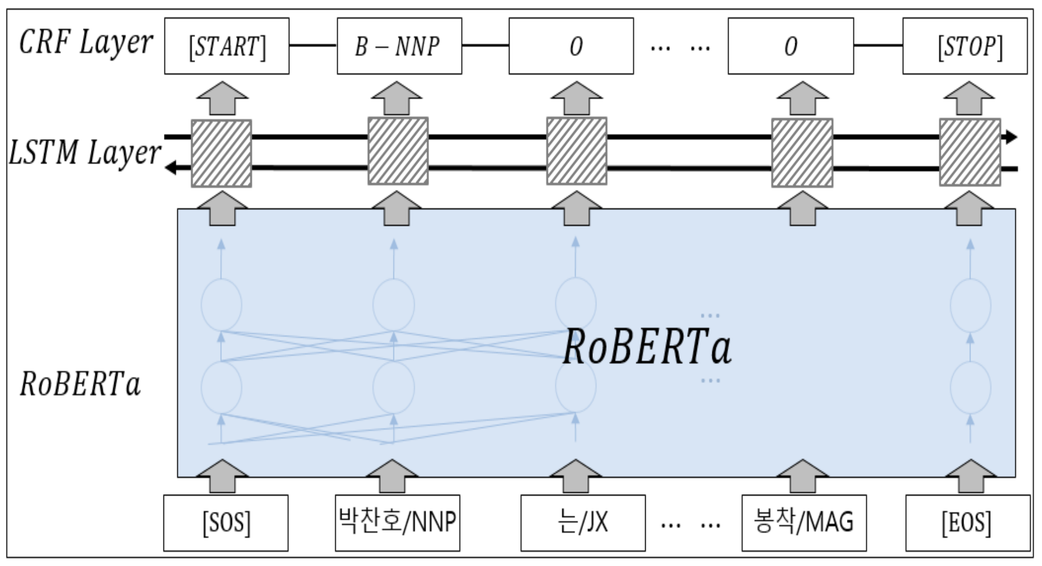
ⅱ. 한국어 RoBERTa 모델

RoBERTa 모델은 기존의 BERT의 학습 과정에서 최적화되지 않은 부분을 최적화 하여 효율적으로 BERT 모델을 학습 할 수 있도록 확장

- **Dynamic Mask LM** 기존의 BERT 모델은 미리 정해진 Mask 되는 단어를 학습하는 모델로 RoBERTa에서는 BERT와 달리 매 학습마다 마스킹하는 단어를 다르게 하는 Dynamic Masking 방식을 사용
- **NSP 테스크 제외** RoBERTa에서는 다음 문장을 예측하는 NSP 태스크에 대해서는 학습하지 않고 하나의 문서에서 문장이 다 채워지지 않으면 다른 문서의 문장으로 채워 넣어 문서 길이가 미리 설정한 최대 토큰 길이 512에 가깝도록 하여 학습 효율을 높임
- **사전 학습** 15GB의 위키피디아 및 뉴스 데이터를 전이기반 형태소 분석 모델로 문서를 형태소 분석하여 사전학습을 위한 학습 코퍼스로 구성. multi-gpu 방식으로 24GB의 용량을 지닌 TITAX RTX 4대를 이용하였으며 총 사전 학습 시간은 2주 가량 소요

Ⅲ. 한국어 자연어 처리 태스크

- **개체명 인식** 개체명 인식에 대한 데이터로는 [10]에서의 ETRI 엑소브레인 언어분석 말뭉치를 사용한다. 그림 1에서 보듯이 “[SOS] 문장 [EOS]”의 토큰 시퀀스를 RoBERTa의 입력으로 사용하고 모델의 출력을 양방향 LSTM을 통해 인코딩 후 출력층에서 CRF와 결합하여 개체명 인식을 수행



- **감성분석** 네이버 한국어 영화 리뷰 데이터 셋을 사용하며 각 영화 리뷰에 대해서 Positive(긍정)/Negative(부정)의 레이블이 부착되어 있음. 개체명 인식과 동일하게 “[SOS] 문장 [EOS]”를 RoBERTa 모델의 입력으로 사용하며 출력에 대해서 양방향 LSTM을 거쳐 얻어진 히든 상태의 평균을 취해 문맥 벡터를 구한 후 출력층에서 해당 문장의 긍/부정 여부를 판별
- **의존파싱** 세종 의존 구문 분석 데이터 셋을 사용하며 RoBERTa 모델의 출력을 얻고 파싱의 단위가 되는 어절을 표상을 위해 어절 내의 형태소+BPE의 마지막 토큰을 취해 단어 표상을 구성 한 후 그래프 기반 방식인 Deep Biaffine 모델을 적용

Ⅳ. 실험 결과

• 개체명 인식 실험 결과

모델	F1
LSTM-CRF	86.53%
BERT(Multilingual)	91.92%
BERT(형태소-태그)	91.58%
RoBERTa[본 연구]	94.79%

• 감성 분석 실험 결과

모델	정확도
LSTM-CRF	79.79%
BERT(Multilingual)	87.43%
BERT(형태소-태그)	86.57%
RoBERTa[본 연구]	89.88%

• 의존 파싱 실험 결과

모델	UAS	LAS
Deep Biaffine	91.78%	89.76%
BERT(형태소-태그	93.24%	92.67%
KorBERT+Biaffine	94.06%	92.00%
RoBERTa + Biaffine[본 연구]	94.32%	92.40%
[Bert] Deep Biaffine	94.42%	92.52%

Ⅴ. 결론 및 향후 연구

본 연구에서는 기존의 BERT 모델을 개선하여 RoBERTa 모델과 미등록어에 강건한 하이브리드 토크나이저를 이용하여 한국어 코퍼스에서 사전학습 한 후 각 응용테스트에 적용하여 기존의 BERT 모델을 사용한 모델보다 더 높은 성능 향상을 보임.