BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정

Cognitive Computing Lab.

박광현[1], 나승훈[1], 신종훈[2], 김영길[2]
[1]전북대학교, [2]한국전자통신연구원
khpark231@gmail.com, nash@jbnu.ac.kr, jhshin82@etri.re.kr, kimyk@etri.re.kr

Ⅰ. BERT

- Attention을 이용한 인코더인 Transformer 기반의 사전 학습 언어모델
- Masking 단어 예측과 다음 문장이 적절한지 판단하는 2가지 문제로 사전 학습을 진행

- Masked Language Model

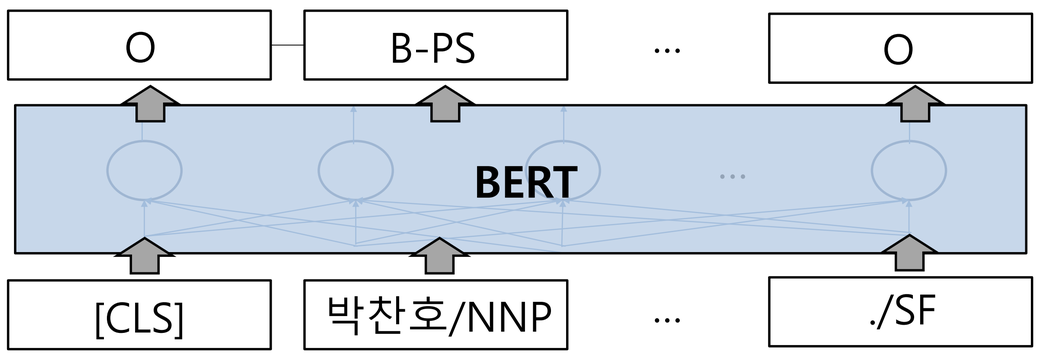
"철수는 밥을 [MASK] ."

- Next Sentence Prediction

"철수는 밥을 먹었다. [SEP] 그리고 학교에 갔다."

- 540MB 정도의 위키피디아 코퍼스를 사용하여 사전학습 진행
- 입력 문장으로 형태소-태그 사용(1997/SN 년/NNB)
- 최대 문장 길이를 128로 학습을 진행한 후 384로 추가 학습 진행
- 인코더 블록 개수 12, 어텐션 헤드 개수 12, 히든 사이즈 768, 드랍아웃 0.1 사용

Ⅱ. 개체명 인식



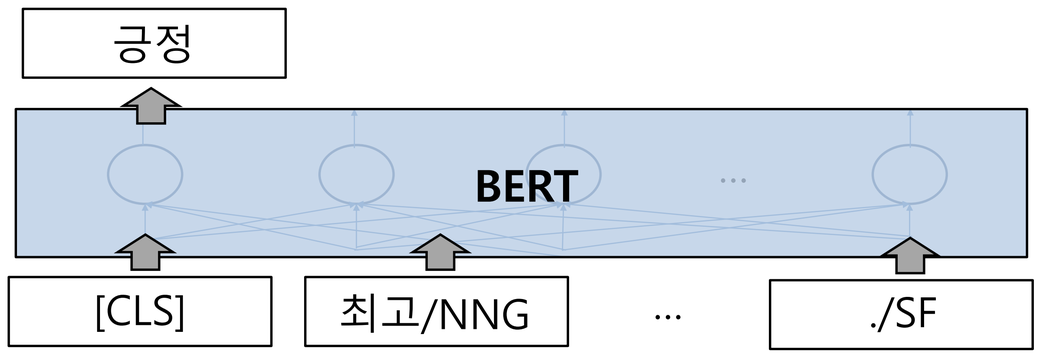
개체명 인식 모델 구조

- 사람, 시간, 날짜, 장소 등 특정한 의미를 가지고 있는 단어를 인식하는 sequence labeling 문제
- ETRI의 엑소브레인 언어분석 말뭉치를 이용하여 실험

모델	F1
LSTM-CRF	86.53%
LSTM-CRF+사전자질	89.34%
BERT(Multilingual)	91.92%
BERT(형태소-태그)	**91.58%**

개체명 인식 실험결과

Ⅲ. 감성분석



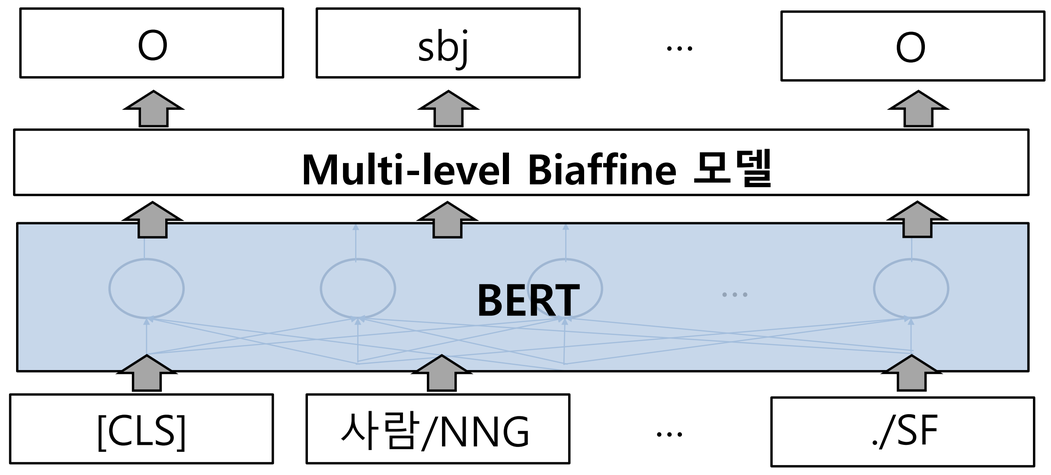
감성분석 모델 구조

- 문장으로부터 좋음, 나쁨 등의 감정을 분석하는 문제
- 네이버 영화리뷰 데이터를 이용하여 실험

모델	F1
LSTM	79.79%
CNN	78.34%
BERT(Multilingual)	87.43%
BERT(형태소-태그)	**86.57%**

감성분석 실험결과

Ⅳ. 의존 파싱



의존파싱 모델 구조

- 문장 내 구성 성분 간의 관계를 분석하여 문장의 구조를 결정하는 문제
- SPMRL 데이터, 세종 데이터를 이용하여 실험
- 평가 척도로 UAS(Unlabeled Attachment Score)와 LAS(Labeled Attachment Score)를 사용

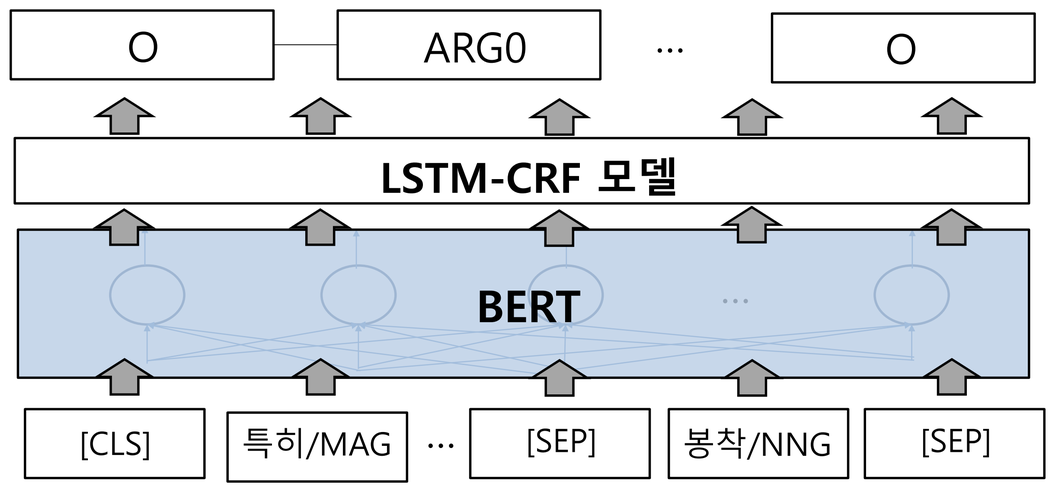
모델	UAS	LAS
Multi-level Biaffine	91.95%	91.38%
Multi-level Biaffine +BERT(Multilingual)	91.93%	91.31%
Deep Biaffine	90.85%	89.31%
BERT(형태소-태그)	**93.24%**	**92.67%**

의존 파싱 SPMRL 데이터 실험결과

모델	UAS	LAS
Multi-level Biaffine	91.84%	89.43%
Multi-level Biaffine +BERT(Multilingual)	92.32%	90.03%
멀티 태스크 포인터 네트워크	91.65%	89.34%
BERT(형태소-태그)	**92.76%**	**90.58%**

의존 파싱 세종 데이터 실험결과

Ⅴ. 의미역 결정



의미역 결정 모델 구조

- 의미역은 서술어에 의해 기술되는 행위나 사태에 대한 명사의 의미 역할을 나타냄
- 서술어에 의해 기술되는 명사구를 논항 이라고 함

"그는(ARG0) 경찰에게 신분증을(ARG1) 보였다."

- Korean propbank데이터를 이용하여 실험

모델	F1
LN LSTM-CRF	78.10%
Stacked LSTM-CRF	78.57%
LSTM-CRF+BERT(형태소-태그)	**84.46%**

의미역 결정 실험결과(AIC)