BERT를 이용한 전이 기반 한국어 형태소 분석 및 품사 태깅

BERT for Transition-based Korean morphological analysis and POS tagging

저자 (Authors)
민진우, 나승훈, 신종훈, 김영길
Jinwoo Min, Seung-Hoon Na, Jong-Hoon Shin, Young-Kil Kim

출처 (Source)
한국정보과학회 학술발표논문집 , 2019.12, 401-403 (3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09301601

APA Style
민진우, 나승훈, 신종훈, 김영길 (2019). BERT를 이용한 전이 기반 한국어 형태소 분석 및 품사 태깅. 한국정보과학회 학술발표논문집, 401-403.

BERT를 이용한 전이 기반 한국어 형태소 분석 및 품사 태깅

민진우 [O1], 나승훈 [2], 신종훈 [3], 김영길 [4]

[12] 전북대학교, [34] 한국전자통신연구원

jinwoomin4488@gmail.com, nash@jbnu.ac.kr, jhshin82@etri.re.kr, kimyk@etri.re.kr

BERT for Transition-based Korean morphological analysis and POS tagging

Jinwoo Min [O1], Seung-Hoon Na[2], Jong-Hoon Shin[3], Young-Kil Kim[4]

[12]Jeonbuk National University, [34]ETRI

요 약

한국어 형태소 분석은 입력된 문장 내의 어절들을 지니는 최소의 단위인 형태소로 분리하고 품사를 부착하는 작업을 의미한다. 기존 한국어 형태소 분석 방법은 음절 기반 연구가 주를 이루고 이를 음절 단위의 순차 태깅 문제로 보고 SVM, CRF혹은 Bi-LSTM-CRF 등을 이용하거나 특정 음절에서 형태소의 경계를 결정하는 전이 기반 모델을 통해 분석하는 모델 등이 연구되었다. 최근 다양한 자연어 처리 분야에서 높은 성능 향상을 보이고 있는 BERT 모델을 한국어 형태소 분석 태스크에 적용하여 BERT를 이용한 전이 기반 형태소 분석 모델을 제안하고 기존의 BERT를 사용하지 않은 모델과의 성능을 비교한다.

1. 서 론

한국어 형태소 분석은 일반적으로 형태소 분석과 품사 태깅의 두가지 과정으로 나뉘며 형태소 분석은 가장 작은 의미를 가진 형태소와 품사 쌍 후보를 생성하는 것이고 품사 태깅은 형태소 분석에서 나온 후보들에서 각 어절의 뜻과 문맥을 고려하여 가장 알맞은 형태소와 품사 쌍을 결정하는 것이다[1]. 딥러닝 기반 형태소 분석[1-3]은 주로 음절 단위 방법으로 연구되었고 입력된 음절열에 대하여 순차 태깅 문제로 보고 품사 태그에 [B,I]등의 태그를 부착한 태그열을 부착하는 Bi-LSTM-CRF[1,3]를 적용하거나 전이 기반 방식[2]을 사용하여 형태소의 끝 음절에서 경계를 결정지음과 동시에 품사를 부착하는 방식이 주로 연구되었다.

BERT[4]는 다층의 양방향 트랜스포머를 인코더로 하여 셀프 어텐션 메커니즘을 통해 모든 레이어에서 전체 문맥 정보를 반영하여 Mask된 단어를 예측하는 Mask LM과 현재 문장의 다음 문장을 예측하는 NSP (Next Sentence Prediction)의 두 태스크로 학습하며 학습된 BERT 모델을 다양한 응용분야에 적용하여 놀라운 성능 향상을 이루었다[4,5]. 본 논문에서는 BERT를 이용한 전이 기반 한국어 형태소 분석 모델을 제안하며 BERT를 적용하지 않은 기존 모델과의 성능을 비교한다.

2. 관련연구

[6]에서 제공되는 미리 학습된 BERT 모델은 영어 단어 단위와 다양한 언어에 적용되기 위해 미리 학습한 BERT-Multilingual 모델이 제공되며 단어 단위 모델은 공백을 단위로 하여 토큰화하는 기본적인 방식이며 Multilingual 모델은 byte-pair-encoding(BPE)[7]이 적용된 subword 단위로 분리된 토큰을 입력으로 하기 때문에 다양한 언어에서 out-of-vocabulary(OOV) 문제에 강건하다[3]. [8]에서는 BERT-Multilingual 모델을 이용하여 입력열을 BPE를 통해 subword 단위로 변환하고 subword 단위에 알맞게 복합태그를 구성하여 이를 LSTM을 통해 복합태그를 예측하고 다시 복합태그를 음절 단위 태그로 분해하는 형태소 분석 모델을 제안하였다.

ETRI에서 공개한 한국어 언어 모델인 KorBERT[9] 역시 형태소, 어절의 두 가지 단위의 입력을 사용한 BERT 모델을 제공하며 형태소 단위 BERT 모델은 각 문장 내의 어절을 형태소 분석기를 이용하여 형태소 단위로 분리한 후 이를 토큰으로 사용하는 모델이며 어절 단위 BERT 모델은 어절을 BPE가 적용된 토큰으로 입력으로 사용하는 모델이다.

3. BERT를 이용한 전이 기반 형태소 분석 모델

본 연구에서는 형태소 분석에 BERT를 적용하기 위해서 미리 학습된 어절 단위 KorBERT를 사용하며 3.1절에서는 KorBERT 모델에 적용하기 위한 입력 포멧에 대하여 설명하고 3.2절에서는 BERT를 이용한 전이 기반 형태소 분석 모델에 대해 설명한다.

3.1 음절 기반 BERT 모델의 입력

어절 단위 KorBERT 모델에 적용하기 위해 [3]과 동일하게 각 어절의 마지막 음절에 "_"를 붙여 음절 시퀀스를 구성하며 시퀀스의 시작과 끝에 각각 [CLS], [SEP]를 추가한다. 형태소 분석을 위한 KorBERT 모델의 입력 예는 다음 표 1과 같다.

표 1. KorBERT 모델의 입력 예

입력 문장
나는 오늘 똑똑히 보았단다.
음절 단위
나, 는, 오, 늘, 똑, 똑, 히, 보, 았, 단, 다, .
KorBERT 모델의 입력
[CLS] 나, 는_, 오, 늘_, 똑, 똑, 히_, 보, 았, 단, 다, ._, [SEP]

별도의 음절 단위로 학습한 BERT 모델이 존재하지 않아 BPE를 적용하여 어절을 토큰화하는 어절단위 KorBERT 모델을 사용하지만 한국어 어절의 경우 BPE를 적용하였을 때 대부분 1음절로 구성된 토큰으로 분리되기 때문에 모델의 단어장에서 대부분의 음절을 포함하여 음절 단위에서 OOV(Out-of-Vocabulary) 문제가 거의 발생하지 않는다.

3.2 BERT를 이용한 전이 기반 한국어 형태소 분석 및 품사 태깅 모델

BERT를 이용한 전이 기반 형태소 분석 모델의 구조는 [2]와 동일하며 차이점은 입력 벡터를 구성할 때 t번째 음절의 음절 임베딩 벡터 $\boldsymbol{c}_t$와 해당 음절이 어절의 시작인지 아닌지를 나타내는지에 대한 [B,I]의 띄어쓰기 임베딩 $\boldsymbol{s}_t$이외에 BERT 모델을 통해 인코딩된 $\boldsymbol{b}_t$ 를 추가적으로 사용한다. 이를 하나로 연결하여 입력열 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$를 구성하며 LSTM을 통해 인코딩하여 은닉열 $\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\}$을 얻는다.

$$\boldsymbol{x}_t = [\boldsymbol{c}_t; \boldsymbol{s}_t; \boldsymbol{b}_t] \quad (1)$$

$$\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\} = LSTM(\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}) \quad (2)$$

[2]와 동일하게 전이 액션을 결정하기 위해 은닉열 $\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\}$에서 자질을 추출하게 되는데 버퍼 B와 스택 S 그리고 전이 액션을 통해 예측된 형태소의 태그를 저장하기 위한 스택 P에서 버퍼는 top 2개, 두 스택는 Top 4개를 취하여 이를 각 자료 구조에 대한 상태표상 $\boldsymbol{B}_t, \boldsymbol{S}_t, \boldsymbol{P}_t$로 표현한다.

$$\boldsymbol{T}_t = Relu(\boldsymbol{W} \cdot [\boldsymbol{B}_t, \boldsymbol{S}_t, \boldsymbol{P}_t]) \quad (3)$$

각 상태표상을 모두 연결한 후 다음 수식 (3)과 같이 연결한 후 비선형 변환을 통해 태거 상태 표상을 $\boldsymbol{T}_t$를 얻는 후 태거 상태표상 $\boldsymbol{T}_t$는 출력층으로 연결되어 다음 전이 액션을 결정한다.

4 실험

본 논문에서 제안한 BERT를 이용한 전이 기반 형태소 분석 모델을 평가하기 위해 세종 형태소 분석 말뭉치를 사용하였다. 학습 셋 202,508 문장과 평가 셋 52,781 문장으로 구성되어 있으며 학습 셋의 5000문장을 별도로 나누어 개발셋으로 사용하였다. 평가 지표로는 형태소 단위 F1과 어절 정확도를 제시한다.

베이스 라인으로 전이 기반 모델 이외에 [3]과 동일한 방식으로 Bi-LSTM-CRF 모델과 이에 BERT를 적용한 두 모델을 제시하며 이는 3.2절의 은닉 표상 $\boldsymbol{h}_i$에 대해 MLP를 적용 한 후 출력 층에서 태그간의 의존성을 모델링하는 CRF와 결합한 모델이다. 배치 사이즈를 64로 하여 전이 기반 모델과 Bi-LSTM-CRF 모델의 형태소 분석 수행 속도를 비교하였을 때 전이 기반 방식이 약 1.4배 빠른 결과를 보였다.

추가로 [8]의 subword 단위 BERT 기반 한국어 형태소 분석 모델을 제시한다. 다음의 표 2는 형태소 분석 실험 결과를 보여주고 있다.

표 2. 형태소 분석 실험 결과

모델	형태소 F1	어절 정확도
CRF [10]	97.60%	96.14%
Phrase-Based-CRF[11]	97.74%	96.35%
전이기반[2]	97.91%	96.65%
subword BERT + LSTM[8]	95.22%	93.90%
전이기반(re-impl)	98.01%	96.78%
Bi-LSTM-CRF(impl)	**98.03%**	**96.81%**
전이기반(re-impl) + BERT	98.01%	96.72%
Bi-LSTM-CRF + BERT	98.01%	96.75%

(위의 모델은 모두 평가 셋이 동일)

표2에서 보듯이 제안한 음절 단위 BERT를 이용한 전이기반 모델이 subword 단위 BERT[8] 적용의 단점을 보완하여 높은 성능을 보이지만 전이 기반 모델과 Bi-LSTM-CRF 모델 모두 BERT를 적용한 모델이 그렇지 않은 모델보다 낮은 성능을 보이고 있다. 본 연구와 동일한 방식으로 BERT를 적용한 [3]의 음절 기반 BERT 임베딩을 사용한 Bi-LSTM-CRF 모델이 동일 셋에서 기존의 성능을 형태소 단위 F1에서 0.72% 가량 성능 향상을 보인 것에 비해 성능이 소폭 하락하여 이에 대한 원인 분석이 필요하다.

평가 셋에서 평가 이전 개발 셋에서의 평가 성능은 BERT를 사용한 모델이 사용하지 않은 모델에서 전이

기반 모델과 Bi-LSTM-CRF 모델 모두 F1 점수 약 0.1%의 성능 향상을 보여 개발 셋에서의 과적합 혹은 개발 셋의 데이터 편향성 문제인지 등을 확인할 예정이다.

5 결론

본 연구에서는 전이 기반 모델에 음절 단위 BERT를 적용하여 실험 결과를 얻었고 전반적으로 성능에 큰 변화가 없었고 어절 단위 KorBERT 모델은 근본적으로 음절 단위 형태소 분석에 적합한 음절 단위로 학습한 BERT가 아니기 때문에 BERT 모델을 적용했을 때의 형태소 분석 성능 향상이 미미하였다.

향후 연구로는 음절 단위 BERT를 포함하여 향상된 모델인 XLNet, 후속 모델들에 대해 음절 정보를 효율적으로 학습할 수 있는 모델에서 학습 후 형태소 분석에 적용할 예정이다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)

참고문헌

[1] 김혜민, 윤정민, 안재현, 배경만, 고영중. 품사 분포와 Bidirectional LSTM-CRFs를 이용한 음절 단위 형태소 분석기, HCL 2016

[2] 민진우, 나승훈, 동적 오라클을 이용한 뉴럴 전이기반 한국어 형태소 분석 및 품사 태깅, HCLT 2018.

[3] 박천음, 이창기, BERT 기반 LSTM-CRF 모델을 이용한 한국어 형태소 분석 및 품사 태깅, HCLT 2019.1

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] 박광현, 나승훈 , 신종훈, 김영길, "BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정", 한국 정보과학회 학술발표논문집, 2019.6

[6] https://github.com/google-research/bert

[7] Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

[8] 민진우, 나승훈, BERT에 기반한 Subword 단위 한국어 형태소 분석, HCLT 2019.

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

[10] Seung-Hoon Na. Conditional Random Fields for Korean Morpheme Segmentation and POS Tagging. ACM Transactions on Asian and Low-Resource. Laungauge Information Processing, 14(3), 2015

[11] Na, Seung-Hoon, and Young-Kil Kim. "Phrase-based statistical model for korean morpheme segmentation and POS tagging." IEICE Transactions on Information and Systems 101.2 (2018)