BERT에 기반한 Subword 단위 한국어 형태소 분석

민진우 [O†], 나승훈[†], 신종훈[††], 김영길[††]

전북대학교[†], ETRI[††]

Jinwoomin4488@gmail.com, nash@jbnu.ac.kr, jhsin82@etri.re.kr, kimyk@etri.re.kr

BERT with subword units for Korean Morphological Analysis

Jin-Woo Min[O†], Seung-Hoon Na[†], Jong-Hun Sin[††], Young-Kil Kim[††]

Jeonbuk National University[†], ETRI[††]

요 약

한국어 형태소 분석은 입력된 문장 내의 어절들을 지니는 최소의 단위인 형태소로 분리하고 품사 부착하는 작업을 의미한다. 기존 한국어 형태소 분석 방법은 음절 기반 연구가 주를 이루고 이를 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 자연어 처리 연구에서 대용량 코퍼스로부터 문맥을 고려한 BERT 등의 언어 모델을 활용한 연구가 각광받고 있다. 본 논문에서는 음절 단위가 아닌 BERT를 이용한 Sub-word 기반 형태소 분석 방법을 제안하고 기분석 사전을 통해 분석하는 과정을 거쳐 세종 한국어 형태소 분석 데이터 셋에서 형태소 단위 F1 : 95.22%, 어절 정확도 : 93.90%의 성능을 얻었다.

주제어: 형태소 분석, LSTM, BERT

1. 서론`

한국어 형태소 분석은 입력된 문장 내의 어절들을 뜻을 지니는 최소의 단위인 형태소로 분리하고 품사를 부착하는 작업을 의미한다. 형태소 분석의 부정확한 분석 결과는 구문 분석, 의미역 결정, 질의 응답 등에 치명적인 영향을 미칠 수 있어 올바른 형태소 분석이 매우 중요하다[4].

기존 한국어 형태소 분석 방법은 주로 음절 기반 연구[4-7]가 주를 이루었는데 입력된 문장을 음절 단위로 하여 순차 레이블링 문제로 보고 CRF[3], SVM[4] 등의 모델을 적용하거나 Bi-LSTM-CRF를 적용하는 방법 전이 기반 방식을 이용하여 형태소의 끝 음절에 품사를 부착하여 형태소를 분석하는 방법이 주를 이루었다.

최근 다양한 자연어 처리 연구서는 대용량 코퍼스로부터 문맥을 고려하여 학습하는 ELMo[2], BERT[1], XLNet[3] 등을 실제 응용 테스크에 파라미터를 미세조정(fine-tuning)하는 방법을 적용하여 큰 성능 향상을 이루었다.

본 연구에서는 BERT 기반 sub-word 단위 Bi-LSTM 한국어 형태소 분석 모델을 적용하여 세종 한국어 형태소 분석 데이터 셋에서 형태소 단위 F1 : 95.22%, 어절 정확도 : 93.90%의 성능을 얻었다.

2. 관련 연구

한국어 형태소 분석은 음절 단위 형태소 분석 방법이 주를 이루었는데 음절 단위의 순차 태깅 문제로 보고 [B(Begin), I(Inside)], 혹은 [B, I, E, S] 등의 태그가 포함된 음절 단위 품사태그를 결정하는 방법이다. 이러한 순차 태깅 문제에 모델은 활용되었다.

[7]에서는 전이 기반 방식을 이용하여 한국어 형태소 분석에 맞는 액션을 정의하고 문장 내의 각 형태소의 경계를 결정하고 품사를 부여하는 방법을 적용하여 CRF 기반 방법보다 높은 성능을 보였다. 이러한 방식은 복합 형태소 방식으로 진행되어 기분석 사전을 통한 원형 복원 및 단위 형태소 분해과정을 거치는데 [12]에서는 기분석 사전이 아닌 경계가 결정된 복합 형태소를 기분석 사전을 이용한 후처리가 아닌 어텐션 기반 Sequence-To-Sequence 모델[13]을 통해 단위 형태소를 생성하는 방식으로 단위 형태소로 분해하는 모델을 제안했다.

최근 자연어 처리 연구는 대용량 코퍼스로부터 양방향 문맥 모델을 Transformer[9]로 학습하는 BERT 모델을 통해 학습하고 학습된 모델을 응용 분야에 활용하는 방법이 각광받고 있다[1]. [10]에서는 의미역 결정. 이에 한국어 BERT 모델을 학습하고 구문 분석, 에서는 BERT 모델을 의미역 결정, 한국어 기계 독해 등에 적용하여 성능을 향상시켰다.

BERT 모델을 순차 태깅 문제에 적용한 [11] 연구에서는 중국어 단어 분할 문제에 순환 구조를 가져 속도가 느린 RNN 기반 인코더가 아닌 상대적으로 적은 파라미터를 갖으며 대용량 코퍼스로부터 사전 학습한(pretrained) 문자 기반 BERT모델과 CRF를 결합하여 Bi-LSTM-CRF 모델에 비해 높은 성능 뿐 아니라 분석 속도를 높이는 연구를 진행하였다.

3. 모델

3.1 Sub word 단위 형태소 분석

Sub word 기반 형태소 분석은 미리 학습된 BERT multi-lingual 모델을 이용한다. 해당 BERT 모델에 적용하기 위해서 입력 문장을 모델의 토크나이저를 이용하여 sub-word 단위로 토큰화하고 음절 단위 태그를 sub-word 단위의 복합태그로 합성하여 출력하는 형식으로 진행하며

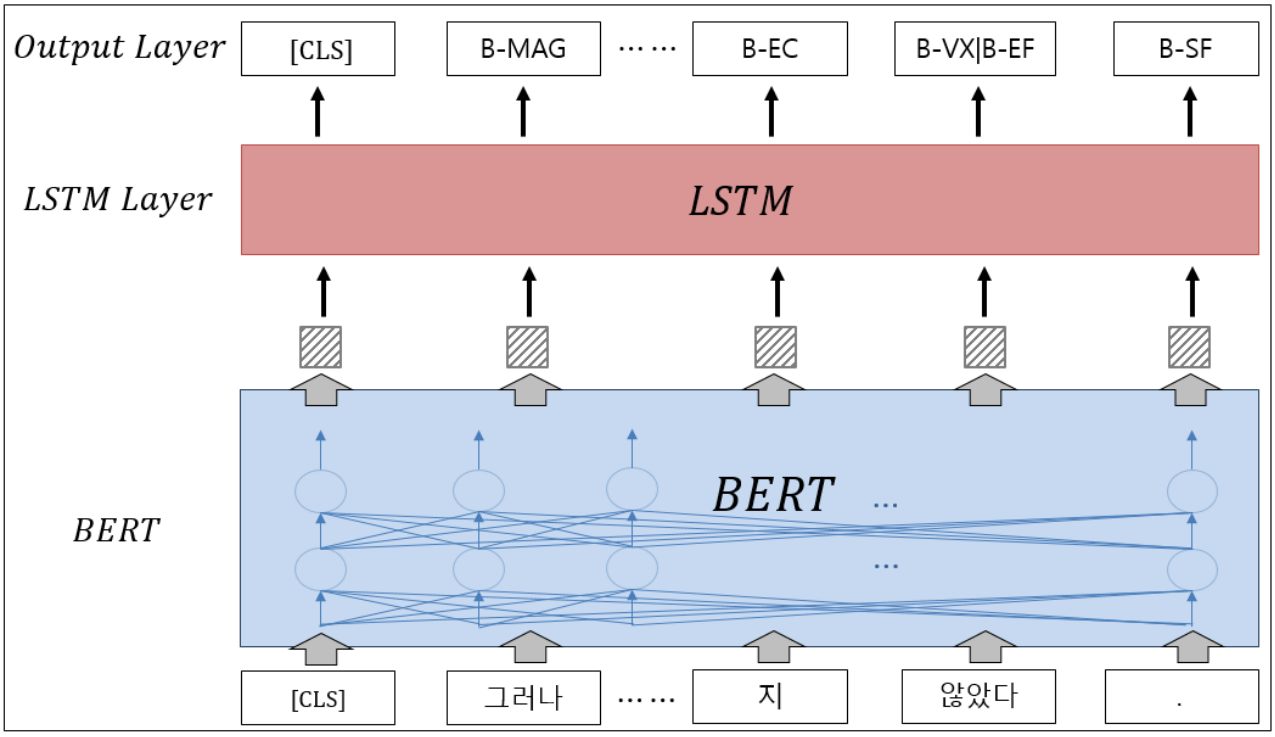

그림 1. BERT 기반 sub-word 단위 형태소 분석 모델

음절 단위 태그를 sub-word 단위 태그로의 합성 규칙은 다음과 같다.

1) 토큰화된 sub-word가 단일 음절이면 해당 음절의 태그를 그대로 사용
2) 토큰화된 sub-word가 복합 음절로 구성되지만 태그가 같으면 단일 태그 사용
3) 토큰화된 sub-word가 복합음절이면서 여러 태그로 구성되면 시작 음절의 태그와 끝 음절의 태그의 결합으로 구성

위의 규칙으로 설명한 Sub-word 복합태그를 구성에 대한 예는 표 1로 보여준다. 위의 예제에서 예제 ①은 규칙 1)에 해당하고 예제 ②, ③은 규칙 2에 해당하며 예제 ④는 규칙 3에 해당한다. 예제 3의 "스트"는 실제 형태소 토스트(NNG)에서 "토"와"스트"가 토큰화 과정에서 분리된 형태로 이 경우에는 단일태그를 부착한다.

표 1. 복합 태그의 구성 예

예제		
①	sub-word	꿈 [B-NNG]
	복합 태그	B-NNG
②	sub-word	하지만 [B-MAG,I-MAG,I-MAG]
	복합 태그	B-MAJ
③	sub-word	스트 [I-NNG I-NNG]
	복합 태그	I-NNG
④	sub-word	에서는[B-JKB,I-JKB,B-JX]
	복합 태그	B-JKB|B-JX

3.2 BERT 기반 Sub word 단위 형태소 분석 모델

본 논문에서 제안한 BERT 기반 Sub word 단위 형태소 분석 모델의 구조는 그림 1과 같다. 그림 1에서 보듯이 sub-word 단위로 토큰화된 문장이 입력 열이 BERT의 입력이 된다. 여기서 BERT 모델은 [1]에서 미리 학습한 다양한 언어를 지원하기 위해 미리 학습된 BERT-Base Multilingual 모델과 토크나이저을 사용한다.

먼저 sub-word 단위로 토큰화된 입력 열을 BERT 모델을 거친 출력을 해당 Sub-word가 어절의 시작인지 아닌지를 나타내는 띄어쓰기 태그[B,I] 임베딩과 결합하여 Bi-LSTM 통해 한번 더 인코딩 한 후 출력 층으로 연결되어 각 sub-word에 대해 확률이 최대가 되는 복합태그를 결정한다.

3.3 sub-word 복합형태소 분해

현재 모델을 통해 분석된 분석 결과는 sub-word 단위로 음절 단위 태그로 변환하는 과정을 거친다. Sub-word 단위 태그를 음절태그로 변환하는 규칙은 다음과 같다.

1) 토큰화된 sub-word가 단일 음절이면 해당 태그가 해당 음절의 태그로 복원
2) 토큰화된 sub-word가 2음절로 구성될 때 단일 태그이면 해당 품사를 그대로 부착하고 복합태그이면 복합태그의 시작태그와 끝 태그를 첫 음절과 끝 음절에 각각 부착
3) 3음절 이상이며 단일 태그가 아닌 경우 기분석 사전에 의한 복원
4) 3)의 기분석 사전에 존재하지 않는 경우는 휴리스틱하게 복원

3)의 기분석 사전은 해당 sub-word가 3음절 이상일 경우에 해당 하며 sub-word의 음절 개수와 해당 복합 태그의 튜플을 키(key)로 하고 음절 단위 형태소의 리스트 해당 키에 해당하는 음절 단위 품사태그의 리스트의 빈도수가 가장 높은 패턴을 값(value)로 하는 기분석 사전을 구성하고 모델의 출력 sub word에 대해 가장 높은 패턴의 단위 형태소 리스트를 선택한다. 4)의 기분석 사전에 해당하지 않는 경우는 sub-word의 길이가 긴 경우에 자주 발생하며 이 경우 형태소 분석 오류가 발생하는 주 원인이 된다.

4. 실험

4.1 실험 세팅

본 논문에서 제안한 모델을 평가하기 위해 [6-8]과 동일한 42개의 품사태그로 구성된 집합 세종 품사 말뭉치를 사용하였으며 음절 단위 복합 형태소 단위로 품사 태깅을 수행하며 총 음절 단위 복합 형태소의 개수는 98개이다. 평가 지표로는 복합 형태소 단위 F1과 어절 정확도로 제시한다.

논문에서 사용한 BERT 모델은 양방향의 Transformer 인코더 블록의 개수 12개, 히든 사이즈 768, 활성함수 gelu, 드랍아웃 0.1 최대 문장 길이 512로 구성되어 있으며 BERT에 대한 학습률은 $5e^{-5}$로 설정하였다. BERT와 연결되는 양방향의 LSTM은 4개의 층으로 구성되며 각 LSTM의 히든 사이즈는 모두 512이며 Optimizer는 adam을 사용했으며 학습률은 0.001, 드랍아웃은 0.33으로 설정하였다.

4.2 실험 결과

본 논문에서 제안한 Bert 기반 Sub-word 단위 Bi-LSTM-CRF 형태소 분석 모델에 대한 베이스 라인으로는 Bert를 사용하지 않은 단순 LSTM 모델을 제시하며 해당 모델은 sub-word 단위의 100차원의 임베딩을 입력으로 사용하였다. 또한, 기계학습 모델인 CRF, Phrase-based CRF, 딥러닝 모델인 Bi-LSTM-CRF, 전이 기반 모델 등의 기존의 연구 결과을 함께 제시한다. 아래의 표 2는 sub-word 단위 형태소 분석 결과를 보여주고 있다.

표 2. 형태소 분석 실험 결과

	형태소 F1	어절 정확도
CRF[3]	97.60%	96.14%
Phrase-Based CRF[4]	97.74%	96.35%
Bi-LSTM-CRF[12]	96.96%	N/A
전이기반 모델[12]	**97.91%**	**96.65%**
sub-word Bi-LSTM	94.38%	92.71%
BERT sub-word Bi-LSTM	95.22%	93.90%

(표에 제시된 모델은 모두 평가셋이 동일)

본 논문에서 제안한 BERT 기반 sub-word 단위 Bi-LSTM 모델이 BERT 인코더를 사용하지 않은 모델 보다 Bert-모델이 어절 기존 음절 단위. 기존의 음절 단위 형태소 분석에서 가장 높은 성을 보이고 있는 전이 기반 모델에 비해 낮은 성능을 보여주고 있다. 2%가량 낮은 성능을 보이고 있다.

연구에서 제안한 모델은 복합태그가 총 1,100 여개로 기존의 태그인 98개보다 10배 이상 많아 Bi-LSTM에 CRF와 결합하기에 메모리 리소스와 시간 복잡도에서 한계가 존재한다. 이를 해결하기 위해서 한국어 형태소 분석에 미리 학습된 한국어 음절 단위의 BERT 모델을 이용하거나 이를 학습하여 활용하는 등의 방법이 필요하다.

또한, google의 subword 단위 multilingual 모델은 소규모로 제한된 한국어 코퍼스로 학습한 모델로 구문 분석에 적용함에 있어 한국어 BERT를 사용한 모델보다 성능 향상 폭이 낮음을 보였다[14].

5. 결론

본 논문에서는 BERT에 기반한 sub-word 단위 형태소 분석 모델을 제안하고 세종 품사 태그 셋에서 형태소 단위 F1 : 95.22%, 어절 정확도 : 93.90%의 성능을 얻었다. 향후 연구로 현재 대용량 한국어 코퍼스로부터 Sub-word BERT 모델을 학습하여 sub-word 단위 형태소 분석 성능 향상의 폭을 높이고 또한, 음절 기반 형태소 분석을 위하여 음절 단위 BERT 모델을 Bi-LSTM-CRF와 전이 기반 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] 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.

[2] Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).

[3] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237.

[4] 김혜민, 윤정민, 안재현, 배경만, 고영중. 품사 분포와 Bidirectional LSTM-CRFs를 이용한 음절단위 형태소 분석기, HCLT 2016

[5] 이창기, “Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델.” 정보과학회논문지: 소프트웨어 및 응용 40.12 (2013): 826-832.

[6] 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
[7] 민진우, 나승훈, 신종훈, 김영길, 동적 오라클을 이용한 뉴럴 전이기반 한국어 형태소 분석 및 품사 태깅, HCLT 2018
[8] 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)
[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
[10] 박광현, 나승훈 , 신종훈, 김영길, "BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정", 한국 정보과학회 학술발표논문집, 2019.6
[11] Huang, W., Cheng, X., Chen, K., Wang, T., & Chu, W. (2019). Toward Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning. arXiv preprint arXiv:1903.04190.
[12] 민진우, 나승훈 , 신종훈, 김영길. (2019). End-to-End 뉴럴 전이 기반 한국어 형태소 분석. 한국정보과학회 학술발표논문집, (), 566-568.
[13] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).
[14] BERT와 ELMo 문맥화 단어 임베딩을 이용한 한국어 의존 파싱, 홍승연, 나승훈, 신종훈, 김영길, KCC 2019