BERT기반 LSTM-CRF 모델을 이용한 한국어 형태소 분석 및 품사 태깅

박천음*ㅇ, 이창기*, 김현기**
강원대학교*, 한국전자통신연구원**
{parkce, leeck}@kangwon.ac.kr, hkk@etri.re.kr

Korean Morphological Analysis and Part-Of-Speech Tagging with LSTM-CRF based on BERT

Cheoneum Park*, Changki Lee*, Hyunki Kim**
Kangwon National University*, Electronics and Telecommunications Research institute**

요 약

기존 딥 러닝을 이용한 형태소 분석 및 품사 태깅(Part-Of-Speech tagging)은 feed-forward neural network에 CRF를 결합하는 방법이나 sequence-to-sequence 모델을 이용한 방법 등의 다양한 모델들이 연구되었다. 본 논문에서는 한국어 형태소 분석 및 품사 태깅을 수행하기 위하여 최근 자연어처리 태스크에서 많은 성능 향상을 보이고 있는 BERT를 기반으로 한 음절 단위 LSTM-CRF 모델을 제안한다. BERT는 양방향성을 가진 트랜스포머(transformer) 인코더를 기반으로 언어 모델을 사전 학습한 것이며, 본 논문에서는 한국어 대용량 코퍼스를 어절 단위로 사전 학습한 KorBERT를 사용한다. 실험 결과, 본 논문에서 제안한 모델이 기존 한국어 형태소 분석 및 품사 태깅 연구들 보다 좋은 (세종 코퍼스) F1 98.74%의 성능을 보였다.

주제어: 형태소 분석, 품사 태깅, BERT, LSTM-CRF

1. 서 론

형태소 분석은 문장에서 최소 의미 단위인 형태소를 구분하는 것이고, 품사 태깅(Part-Of-Speech tagging)은 형태소의 여러 가능한 품사들 중 적절한 품사로 분류하는 것이다. 한국어는 교착어에 속하기 때문에 대부분의 한국어 자연어처리 태스크(의존구문분석, 개체명인식, 의미역결정, 상호참조해결 등) [1, 2, 3, 4,]는 형태소 분석과 품사 태깅 결과를 기반으로 한다.

기존 한국어 형태소 분석은 주어진 입력 문장에서 가능한 형태소 후보들을 모두 생성하고 품사 태깅 모델을 이용하여 최적의 결과를 결정한다. 이때 기분석 사전이나, 원형 복원 등 다양한 언어 자원이 필요하며, 형태소 분석 및 품사 태깅은 파이프라인(pipeline)으로 진행되기 때문에 오류가 누적되는 문제가 있다. 이러한 문제를 해결하기 위하여 [5, 6, 7, 15]은 Structural SVM이나, feed-forward neural network (FFNN)에 conditional random field (CRF) [8]를 결합하는 방법이나 sequence-to-sequence 모델을 이용한 방법 등을 이용하여 한국어 형태소 분석 및 품사 태깅을 수행하였다.

최근 다양한 자연어처리 태스크에 적용되어 성능 향상에 도움을 주고 있는 Bidirectional Encoder Representations from Transformers (BERT) [9]는 대용량 말뭉치를 언어 모델로 학습한 모델이다. BERT는 양방향성을 가진 트랜스포머(transformer)의 인코더(encoder) [10]를 기반으로 하며, 네트워크의 모든 레이어에서 전체 문맥 정보를 확인하여 언어 모델을 학습한다. BERT는 byte pair encoding (BPE) [11]이 적용된 토큰을 입력으로 하기 때문에 out-of-vocabulary (OOV) 문제에 강하다. 일반적인 BERT는 문장 내에서 임의의 단어에 대하여 마스킹(masking)하고 이를 예측하는 masked language modeling (masked LM)과 다음 문장 예측을 수행한다.

본 논문에서는 한국어 형태소 분석 및 품사 태깅을 수행하기 위하여 BERT 기반 음절 단위 LSTM-CRF 모델을 제안하며, ETRI에서 대용량 한국어 데이터를 어절 단위로 사전 학습한 KorBERT [4]를 사용한다.

2. BERT 기반 음절 단위 LSTM-CRF 모델을 이용한 한국어 형태소 분석 및 품사 태깅

본 논문에서는 사전 학습 된 BERT를 기반으로 음절 기반 LSTM-CRF 모델을 이용하여 한국어 형태소 분석 및 품사 태깅을 수행한다. 제안된 모델은 [그림 1]과 같으며, 주어진 음절 단위 입력열에 대한 BERT 출력에 어절 경계 자질 등을 추가한 후에 bidirectional LSTM (bi-LSTM) [12, 13]의 입력으로 사용하고, 출력 정보 간의 의존성을 모델링 하기 위해 CRF 레이어를 추가한다.

2.1. BERT를 이용한 한국어 형태소 분석 및 품사 태깅의 입력 구조

한국어 형태소 분석 및 품사 태깅을 수행하기 위한 모델의 입력은 음절 단위로 나뉜 문장이며, 출력은 각 입력에 대응되는 품사 태그이다. 이때 형태소의 경계를 구분하기 위하여 출력 품사 태그에 BI 태그(B는 형태소의 시작, I는 형태소가 이어짐을 뜻함)를 적용한다. 본 논문에서 사용하는 KorBERT는 어절 단위에 BPE를 적용하여 만들어진 단어

사전으로 사전 학습된 것이다. KorBERT를 사용하기 위하여 어절의 마지막 음절에 '_'를 붙이며, 입력열의 시작과 끝에 각각 [CLS]와 [SEP] 토큰을 추가한다. 본 논문에서는 어절 구분을 위한 어절 범위 자질을 추가하며, 어절 구분을 위하여 BIE 태그(B는 어절의 시작, I는 어절이 이어짐, E는 어절의 끝을 뜻함)를 사용한다. 형태소 분석 및 품사 태깅의 입출력과 KorBERT 입력, 어절 구분 자질에 대한 예는 [표 1]과 같다.

표 1. 한국어 형태소 분석 및 품사 태깅의 입출력 및 BPE 적용 예제

입력	신문 시장은 여느 때보다 탈법과 불법이 난무하고 있다.
음절 단위 입력 문장	신 문 시 장 은 여 느 때 보 다 탈 법 과 불 법 이 난 무 하 고 있 다 .
Kor BERT 입력	[CLS] 신 문_ 시 장 은_ 여 느_ 때 보 다_ 탈 법 과_ 불 법 이_ 난 무 하 고_ 있 다 ._ [SEP]
어절 범위 자질	E B E B I E B E B I E B I E B I E B I I E B I E E
형태소 분석 및 품사 태깅 출력	B-NNG I-NNG B-NNG I-NNG B-JX B-MM I-MM B-NNG B-JKB I-JKB B-NNG I-NNG B-JC B-NNG I-NNG B-JKS B-NNG I-NNG B-XSV B-EC B-VX B-EF B-SF

2.2. 한국어 형태소 분석 및 품사 태깅을 위한 BERT 기반 LSTM-CRF 모델

한국어 형태소 분석 및 품사 태깅 모델의 학습 데이터는 음절 단위 입력열 $X = \{x_1, x_2, \dots, x_n\}$와 어절 구분 입력 자질 $F = \{f_1, f_2, \dots, f_n\}$, 형태소 분석 및 품사 태깅 결과 $Y = \{y_1, y_2, \dots, y_n\}$로 구성된다. [그림 1]과 같이 입력된 각 음절은 토큰 임베딩(token embedding)을 얻고, 세그먼트 임베딩(segment embedding)과 포지션 임베딩(position embedding)을 더하여 BERT의 입력 표현 e_i를 만든다. 그 후, 사전 학습된 BERT 모델을 거쳐 BERT의 단어 표현 b_i를 얻으며, 식은 다음과 같다.

$$t_i = emb_{token}(x_i) \quad (1)$$
$$s_i = emb_{seg}(seg_i) \quad (1)$$
$$p_i = emb_{pos}(pos_i) \quad (1)$$

$$e_i = t_i + s_i + p_i \quad (2)$$

$$b_i = \text{BERT}(e_i) \quad (3)$$

생성된 b_i를 기반으로 bi-LSTM을 수행하여 히든 스테이트 $\mathbf{r}_i$를 얻으며, 식 (4)와 같다. 식 (4)를 수행하기 전에 식 (5, 6)과 같이 BERT 단어 표현 b_i와 어절 범위 자질 임베딩 h_i^f를 함께 연결하여 bi-LSTM입력으로 넘긴다.

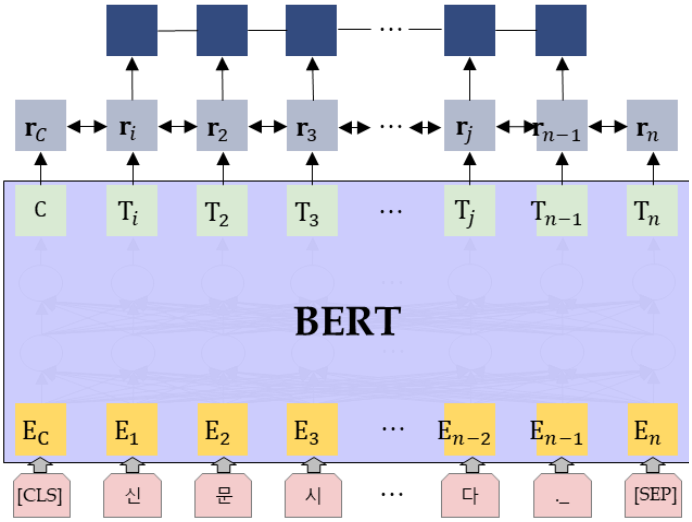

그림 1. BERT 기반 LSTM-CRF를 이용한 한국어 형태소 분석 및 품사 태깅 모델

$$\mathbf{r}_i = biLSTM(h_i) \quad (4)$$

$$h_i = [b_i; h_i^f] \quad (5)$$

$$h_i^f = emb_{feat}(f_i) \quad (6)$$

마지막으로, bi-LSTM으로 인코딩된 히든 스테이트는 output layer인 CRF의 입력으로 주어지고, CRF는 음절 단위 품사 태그를 출력한다. 식은 다음과 같다.

$$s(x, y) = \sum_{t=1}^{T} A(y_{t-1}, y_t) + y_t \quad (7)$$
$$\log p(y|x) = s(x, y) - log \sum_{y'} \exp(s(x, y')) \quad (8)$$

위 식에서 $A(y_{t-1}, y_t)$는 이전 출력 y_{t-1}에서 현재 출력 y_t로 전이될 확률을 계산하는 함수이고 $s(x, y)$는 출력열 y의 점수를 구하는 함수이다.

3. 실험

한국어 형태소 분석 및 품사 태깅을 위하여 세종 코퍼스[14]를 사용하였으며, 학습 데이터는 753,549문장을 사용하고, 평가 데이터는 [15]와 같은 9,379문장을 사용하였다.

본 논문에서는 ETRI에서 사전 학습한 KorBERT 모델[4]을 이용하였다. KorBERT는 웹에서 수집한 뉴스 및 위키피디아 데이터 총 23.5 기가바이트를 사전 학습하였다. KorBERT는 구글의 BERT-base와 같은 하이퍼 파라미터를 사용하며 다음과 같다. 트랜스포머 블록 수는 12, 히든 레이어 차원 수는 768, 어텐션 헤드 수는 12, 히든

레이어의 드랍아웃(dropout)은 0.1, 언어 모델 학습을 위한 최대 문장 길이는 512로 설정됐으며, 각 히든 레이어의 활성 함수는 gelu [16]가 사용된다.

BERT 기반 LSTM-CRF를 이용한 형태소 분석 및 품사 태깅 모델에서 사용한 LSTM의 히든 레이어 차원 수는 BERT의 히든 레이어 차원 수와 같은 768, LSTM의 히든 레이어 스택 수는 1, LSTM의 드랍아웃은 0.1로 설정하였다. 자질 표현의 차원 수는 20으로 설정하였다. 학습 배치 크기는 16으로 하였고, 학습율(learning rate)은 사전 학습 된 BERT를 fine-tuning 하기 때문에 5e-5로 설정하였다. 에폭(epoch) 수는 3으로 설정하고, 학습 알고리즘은 Adam [17]을 사용하며 Adam 가중치 감소(weight decay)는 1e-02로 설정하였다.

[표 1]은 세종 코퍼스에 대하여 본 논문에서 제안한 BERT 기반 LSTM-CRF 모델의 성능과 기존 한국어 형태소 분석 및 품사 태깅 연구[6, 15, 18, 19]들의 성능을 비교한 것이다. [표 2]에서 [6, 18]은 본 논문과 다른 평가 셋을 사용하였다. 실험 결과, 본 논문에서 제안한 모델이 98.74%로 기존 연구들 보다 높은 성능을 보였다.

표 1. 세종 코퍼스 형태소 분석 및 품사 태깅 성능 비교 (*은 평가 셋이 다름)

Models	F1
나승훈[18]: CRF*	97.65
이건일[6]: Sequence-to-sequence*	97.15
이창기[15]: structural SVM	98.03
RNN-search [13]	95.92
황현선[29]: Copying mechanism	97.08
BERT + LSTM-CRF (ours)	**98.74**

4. 결론

본 논문에서는 한국어 대용량 코퍼스로 사전 학습한 BERT (KorBERT)를 기반으로 한 LSTM-CRF 모델을 형태소 분석 및 품사 태깅에 적용하였고, 입력 문장을 음절 단위로 나눈 뒤 어절을 구분하기 위하여 어절 범위 자질 자질을 추가하였다. 실험 결과, 본 논문에서 제안한 모델이 한국어 형태소 분석 및 품사 태깅에서 기존 연구들보다 좋은 성능을 보였다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행 된 연구임(No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능진화형 Wise QA 플랫폼 기술 개발)

참고문헌

[1] 나승훈, et al. Deep Biaffine Attention을 이용한 한국어 의존 파싱, *KCC*, pp. 584-586, 2017.

[2] 이창기, 김준석, 김정희, 김현기. 딥 러닝을 이용한 개체명 인식, *KIISE 동계학술발표회 논문집*, 2014.

[3] 배장성, 이창기, 임수종, 김현기. BERT를 이용한 한국어 의미역 결정, *KCC*, 2019.

[4] 박천음, 김기훈, 이창기, 임준호, 류지희, 김현기. BERT기반 Deep Biaffine을 이용한 한국어 상호참조 해결, *KCC*, 2019

[5] 나승훈, 정상근. 딥 러닝에 기반한 한국어 품사 태깅. *KIISE 동계학술발표회 논문집*, 2014.

[6] 이건일, 이의현, 이종혁. Sequence-to-sequence 기반 한국어 형태소 분석 및 품사 태깅. *JOK, Vol. 44, No. 1*, 2017, pp. 57-62.

[7] 황현선, 이창기. Copying mechanism을 이용한 Sequence-to-Sequence 모델기반 한국어 형태소 분석. *KSC*, 2016.

[8] Dan Roth and Wen-tau Yih. Integer linear programming inference for conditional random fields. *In Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*. 2015, pp. 736–743.

[9] J. Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint *arXiv:1810.04805*, 2018.

[10] A. Vaswani, et al. Attention Is All You Need. *Neural Information Processing Systems (NIPS)*, pp. 5998-6008, 2017.

[11] R. Sennrich, et al. Neural Machine Translation of Rare Words with Subword Units. *In Proc. of ACL*, pp.1715-1725, 2016.

[12] S. Hochreiter and J. Schmidhuber. Long Short-term Memory. Neural Computation, pp. 1735-1780, 1997.

[13] D. Bahdanau, et al. Neural machine translation by jointly learning to align and translate. *Proc. of ICLR' 15*, arXiv:1409.0473, 2015.

[14] 국립국어원. 21세기세종계획. 2012.

[15] 이창기. Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델. *JOK, Vol.40, No. 12*, 2013. pp. 826-832.

[16] D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs). arXiv preprint arXiv:1606.08415v3, 2018.

[17] D.P. Kingma and J.L. Ba. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. arXiv preprint arXiv:1412.6980v9, 2015.

[18] 나승훈, 김영길. 구기반 통계적 모델을 이용한 한국어 형태소 분할 및 품사 태깅. *한국정보과학회 학술발표논문집*, 2014, pp. 571-573.

[19] 황현선, 이창기. Copying mechanism을 이용한 Sequence-to-Sequence 모델기반 한국어 형태소 분석. *한국정보과학회 학술발표논문집*, 2016. pp. 443-445.