Sequence-to-Sequence 와 BERT-LSTM을 활용한 한국어 형태소 분석 및 품사 태깅 파이프라인 모델

윤준영○, 이재성
충북대학교
junyoung292@cbnu.ac.kr, jasonlee@cbnu.ac.kr

A Pipeline Model for Korean Morphological Analysis and Part-of-Speech Tagging Using Sequence-to-Sequence and BERT-LSTM

Jun Young Youn○, Jae Sung Lee
Chungbuk National University

요 약

최근 한국어 형태소 분석 및 품사 태깅에 관한 연구는 주로 표층형에 대해 형태소 분리와 품사 태깅을 먼저하고, 추가 언어자원을 사용하여 후처리로 형태소 원형과 품사를 복원해왔다. 본 연구에서는 형태소 분석 및 품사 태깅을 두 단계로 나누어, Sequence-to-Sequence를 활용하여 형태소 원형 복원을 먼저 하고, 최근 자연어처리의 다양한 분야에서 우수한 성능을 보이는 BERT를 활용하여 형태소 분리 및 품사 태깅을 하였다. 본 논문에서는 두 단계를 파이프라인으로 연결하였고, 제안하는 형태소 분석 및 품사 태깅 파이프라인 모델은 음절 정확도가 98.39%, 형태소 정확도 98.27%, 어절 정확도 96.31%의 성능을 보였다.

주제어: 형태소 분석, 형태소 품사 태깅, Sequence-to-Sequence, BERT, 파이프라인

1. 서론

형태소 분석 및 품사 태깅은 문장에 포함된 어절들에서 의미를 지니는 최소 단위인 형태소를 추출하고, 추출된 형태소에 대한 품사(Part-of-Speech)를 부착하는 작업이다. 교착어에 속하는 한국어는 개체명인식, 의존구문분석, 단어 의미 모호성 해소 등 다양한 자연어처리 연구에서 형태소 분석 결과를 입력으로 사용하기 때문에 정확한 형태소 분석과 품사 태깅이 필요하다[1-3].

형태소 분석 및 품사 태깅 연구는 크게 형태소 원형 복원, 형태소 분리, 형태소 품사 태깅 3단계로 볼 수 있다[4]. 형태소 분석 단계를 파이프라인(pipeline)으로 적용할 경우 이전 단계의 오류가 누적되는 문제가 발생한다. 최근 머신러닝 기반의 한국어 형태소 분석 및 품사 태깅 연구는 주로 문장 단위로 처리하며, 표층형에서 형태소 분리 및 품사 태깅을 먼저 수행하고, 후처리로 사전을 활용하여 형태소 원형 복원을 해왔다[5-7]. 이때 기분석 사전 또는 원형 복원 사전과 같은 추가적인 언어자원을 필요로 한다[8-9]. 본 연구에서는 한국어 형태소 분석 및 품사 태깅을 위해 Sequence-to-Sequence를 활용하여 먼저 형태소 원형을 복원하고, 형태소 원형 복원 결과에 BERT-LSTM을 활용하여 형태소 분리 와 품사 태깅을 동시에 하는 파이프라인 모델을 제안한다.

Sequence-to-Sequence[10] 모델은 가변길이의 시퀀스를 다른 형태의 시퀀스로 변환하는 모델이며, 입력 시퀀스를 축약된 표현으로 출력하는 인코더와 축약표현을 기반으로 다른 문장을 출력하는 디코더로 구성된다. Sequence-to-Sequence는 기계번역, 음성인식 등의 분야에서 주로 사용되는데 [11,12]는 형태소 분석 및 품사 태깅 문제를 시퀀스 번역 문제로 접근하였다. 이와 같은 방식은 Sequence-to-Sequence가 임의의 길이를 갖는 시퀀스를 입력으로 사용할 수 있고, 형태소 분석 및 품사 태깅 과정에서 추가 언어자원에 의존할 필요가 없는 장점이 있다.

BERT 모델은 언어 모델의 하나로 주의(attention)를 기반으로 하는 트랜스포머를 사용하여 대용량 말뭉치를 학습한 모델이다[13]. 사전학습(pre-trained)된 언어 모델을 이용하면 이를 사용하는 자연어 처리 시스템의 성능을 향상시키는데, 정밀 조정(fine-tuning)이나 특성(feature) 추가를 통해 최근 자연어 처리의 다양한 분야에서 우수한 성능을 보이고 있다. 본 연구에서는 ETRI에서 대용량 한국어 말뭉치를 사용하여 어절단위로 학습한 KorBERT를 사전학습 모델로 사용한다.

2. Sequence-to-Sequence 와 BERT-LSTM를 활용한 형태소 분석 및 품사 태깅

본 논문에서는 형태소 분석을 1)형태소 원형 복원과 2)형태소 분리 및 품사 태깅의 두 단계로 나누어 처리한다. 먼저 Sequence-to-Sequence기반의 형태소 원형 복원 단계, 다음으로 BERT-LSTM기반의 형태소 분리 및 품사 태깅 단계 순서로 진행된다. 형태소 원형 복원 모델과 형태소 분리 및 품사 태깅 모델은 학습시 각 단계에서 독립적으로 학습하며 사용한 모델의 자세한 설명은 아래와 같다.

2.1 Sequence-to-Sequence 기반 원형 복원 모델

형태소 원형 복원을 위해 주의(attention) 기반의 Sequence-to-Sequence를 사용하였다[14]. 입력과 출력은 원문과 형태소 원형 복원 결과를 각각 음절 단위로 분해하여 사용하였으며, 문장에 포함된 어절간의 구분을 위해 “<B>” 토큰을 추가하였다. 원형 복원 단계에서 사용된 입력과 출력에 대한 예는 아래의 표 1과 같다.

표 1. Sequence-to-Sequence 기반 형태소 원형 복원 모델의 입력 및 출력

입력(원문)	누 가 <B> 가 르 쳐 <B> 줬 느 냐 고
출력(형태소 원형)	누 구 가 <B> 가 르 치 어 <B> 주 었 느 냐 고

2.2 BERT-LSTM 기반 형태소 분리 및 품사 태깅 모델

형태소 분리 및 품사 태깅을 위해 [5-7]와 같이 음절 단위 품사 태깅을 수행하였다. 음절 단위 품사 태깅은 어절 단위로 사전 학습된(pre-trained)된 KorBERT를 사용하였고 이를 음절 단위로 정밀 조정(fine-tuning) 하였다. 음절 단위 품사 태깅의 입력은 형태소 원형을 KorBERT의 입력으로 변환하여 사용하였으며(어절의 마지막 음절에 “_”를 추가하고 문장 앞뒤로 [CLS]와 [SEP] 토큰을 각각 추가), 출력은 각 입력에 대응되는 분리 태그 BI(B:형태소의 시작, I: 형태소 이어짐을 의미)와 품사 태그를 결합하여 사용하였다. 분리 및 품사 태깅을 위해 정밀조정에 사용한 입력과 출력에 대한 예는 아래의 표 2와 같다.

표 2. 형태소 분리 및 품사 태깅 모델의 입력 및 출력

형태소 원형	누 구 가 <B> 가 르 치 어 <B> 주 었 느 냐 고
KorBERT 입력	[CLS] 누 구 가_ 가 르 치 어_ 주 었 느 냐 고_ [SEP]
출력(형태소 분리 태그 및 품사 태그 결합)	B-NP I-NP B-JKS B-VV I-VV I-VV B-EC B-VX B-EP B-EC I_EC I-EC

2.3 Sequence-to-Sequence 및 BERT-LSTM을 활용한 한국어 형태소 분석 및 품사 태깅 파이프라인 모델

본 논문에서 제안하는 형태소 분석 및 품사 태깅을 위한 Sequence-to-Sequence 및 BERT-LSTM 모델은 학습 시에는 정답 셋을 사용하여 각 단계가 독립적으로 학습되며 평가시에는 아래의 그림 1과 같이 파이프라인으로 처리된다. 형태소 분석용 원시 입력문장을 Sequence-to-Sequence의 입력으로 사용하여 원형 복원을 하고, 원형 복원 결과를 변환(어절을 구분하는 “<B>” 태그를 제거하고, 어절의 마지막 음절에 “_”를 추가)하여 BERT-LSTM의 입력으로 사용한다. 최종적으로 원형 복원 결과와 분리 및 품사 태깅 결과를 사용한다.

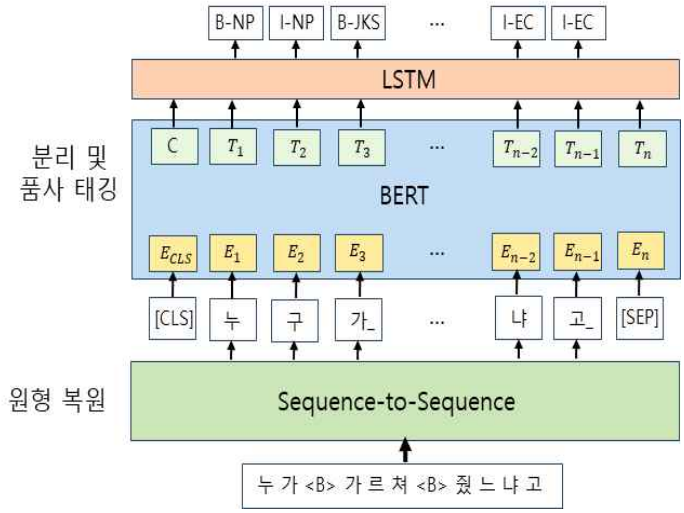

그림 1. Sequence-to-Sequence 및 BERT-LSTM을 활용한 한국어 형태소 분석 및 품사 태깅 파이프라인 모델

3. 실험 및 평가

실험은 세종 말뭉치[15]를 전처리 후 총 66만 문장을 사용하였으며, 그중 90%는 학습데이터로, 10%는 평가데이터로 사용하였다. 전처리로 한글, 영어, 숫자, 일부 특수문자(마침표, 쉼표, 따옴표 등)를 제외한 모든 문자를 제거하였다. 또한 세종 말뭉치에서 원형 문장과 형태소 분석 결과 문장에 포함된 어절의 개수가 다른 경우(약 9만 문장)와 sequence-to-sequence와 BERT의 입력으로 사용하기 어려운 음절 길이 500 이상의 문장과 음절 길이 30 이상을 갖는 형태소를 포함하는 문장(약 500 문장)은 제거하였다.

3.1 Sequence-to-Sequence 형태소 원형 복원 모델

원형 복원 모델 학습을 위해 인코더의 순환신경망의 유닛은 64개, 디코더는 64개, 임베딩 차원은 64로 설정하였다. 학습을 위해 stochastic gradient descent(SGD)를 사용하고 학습률은 0.0001, 배치크기는 10으로 설정하였다. 정답 셋을 기준으로 평가한 원형 복원 모델의 성능은 아래의 표 3과 같다.

표 3. 형태소 원형 복원 모델 성능 비교

모델	음절 정확도	어절 정확도
원형 복원 사전 이용[8]	98.12	96.64
음절 복원 사전 이용[8]	97.99	96.47
sequence-to-sequence 원형 복원 (제안 모델)	99.78	99.26

3.2 BERT-LSTM 형태소 분리 및 품사 태깅 모델

분리 및 품사 태깅 모델에서 사용한 KorBERT 모델은 트랜스포머 블록수 12, 히든 레이어 차원수 768, 최대 문장 길이 512로 구성되어 있으며 히든레이어의 드랍아웃 0.1, 학습률 $5e^{-5}$, 활성화 함수로 gelu가 사용된다. BERT와 연결되는 LSTM은 bi-LSTM을 사용하였으며, bi-LSTM은 하나의 층으로, 히든 레이어 차원수는

KorBERT와 동일한 768, 배치크기 4, 에폭(epoch)은 10으로 설정하였다. 정답 셋을 기준으로 평가한 분리 및 품사 태깅 모델의 성능은 아래의 표 4와 같다.

표 4. 형태소 분리 및 품사 태깅 모델 성능 비교
(입력으로 정답 형태소 원형 사용)

모델	음절 정확도	형태소 정확도	어절 정확도
BERT-LSTM +CRF[5]*	98.74	-	-
Structural SVM[6]*	98.03	-	-
BERT-LSTM (제안 모델)	98.49	98.35	96.44

* 모델은 입력으로 원형 복원 이전의 표층형 데이터(원시문장)를 사용하여 직접 비교가 불가능하므로 참고용으로 표시함

3.3 Sequence-to-Sequence 및 BERT-LSTM 파이프라인 모델

표 5는 본 논문에서 제안하는 파이프라인 모델과 기존 한국어 형태소 분석 및 품사 태깅 연구들의 성능을 비교한 것이다. 파이프라인 모델은 Sequence-to-Sequence 원형 복원 모델의 출력 결과를 BERT-LSTM 모델의 입력으로 사용하여 평가하였다. 실험 결과, 본 논문에서 제안하는 형태소 분석 및 품사 태깅 파이프라인 모델은 음절 정확도 98.39%, 형태소 정확도 98.27%, 어절 정확도 96.31%의 성능을 보였다.

표 5. 형태소 분석 및 품사 태깅 성능 비교
(입력으로 원시 문장 사용)

모델	음절 정확도	형태소 정확도	어절 정확도
CRF[7]	-	97.65	96.24
Sequence-to-Sequence[11]	-	97.15	95.33
Bert sub-word Bi-LSTM[16]	-	95.22	93.90
파이프라인 모델 (제안 모델)	**98.39**	**98.27**	**96.31**

3.4 결과 분석

기존 원형 복원 사전을 구축하여 사용하는 연구들은 원형 복원 사전 구축을 위한 규칙을 필요로 한다[7-9]. 복합 태그를 사용한 경우 출력 태그의 개수가 증가하여 성능 저하의 원인이 될 수 있다.[1] 또한 형태소 원형 복원 사전을 활용하는 경우, 코퍼스로부터 자동으로 사전을 구축하기 때문에 원형 복원의 모호성이 있는 단어를 처리하는데 어려움이 있다[6-7]. 아래의 표 6은 형태소 원형 복원 사전을 활용한 경우 발생하는 오류의 예이다. "향이 나는 커피"라는 문장이 주어질 때, 단어 "나는"의 형태소 분리 및 품사 태깅 결과는 "나/VV+는/ETM"이 되고, "나/VV+는/ETM"의 원형 복원 형태는 "날/VV+는/ETM"과 "나/VV+는/ETM"이 있을 수 있다. 주변 문맥에 따라 실제 원형 복원 결과는 "나/VV+는/ETM"에 해당되지만, 형태소 원형 복원 사전을 사용할 경우 의미와 관계없이 높은 빈도수를 갖는 "날/VV+는/ETM"으로 원형 복원이 이루어진다[6]. 반면 본 연구에서 제안하는 파이프라인 모델은 문맥을 고려하여 형태소 원형 복원을 먼저하고, 형태소 분리 및 품사 태깅을 하였기 때문에 이러한 문제를 완화시키는 것으로 해석된다.

표 6. 형태소 원형 복원 사전을 활용한 오류의 예

모델	형태소 분석 및 품사 태깅 과정
형태소 분리 및 품사 태깅 후 빈도수 기반의 형태소 원형 복원 사전 활용	원시 문장 : 향이 나는 커피 1) 형태소 분리 및 품사 태깅 : - 나/VV+는/ETM 2) 빈도수 기반의 형태소 원형 복원 사전 활용 : - 날/VV+는/ETM
제안 모델 (파이프라인)	원시 문장 : 향이 나는 커피 1) 형태소 원형 복원 : - 나 는 2) 형태소 분리 및 품사 태깅 : - 나/VV+는/ETM

4. 결론

본 논문에서는 Sequence-to-Sequence 기반 형태소 원형 복원과 BERT-LSTM 기반 형태소 분리 및 품사 태깅 모델을 파이프라인으로 처리하는 모델을 제안하고 기존연구들과 비교하였다. 원형 복원 단계의 오류가 다음 단계에 누적이 되는 이유로 표층형에서 분리 태깅 후 원형 복원 사전을 이용하는 연구[5,6,7,9,15]와 다르게 Sequence-to-Sequence를 활용하여 원형 복원 먼저 적용하고, 원형 복원 결과를 BERT-LSTM의 입력으로 사용하여 분리 및 품사 태깅을 적용하였다. 그 결과, 원형 복원 사전을 사용하지 않고 우수한 성능을 보였으며, 특히 형태소 단위 분석 정확도와 어절 단위 분석 정확도는 기존 연구에 비해 상대적으로 우수함을 보였다.

감사의 글

이 (성과물)은 중소벤처기업부 '산업전문인력역량강화사업'의 재원으로 한국산학연협회(AURI)의 지원을 받아 수행된 연구임. (2020년 기업연계형연구개발인력양성

1) [8]에서 기본 태그를 사용한 경우 22개의 태그가 사용된 반면, 복합 태그를 사용한 경우 58개의 태그를 사용한다. (본 연구에서는 36개의 세종 품사 태그를 사용)

사업, 과제번호 : S2929950)

참고문헌

[1] 이창기, et al. “딥 러닝을 이용한 개체명 인식.” 한국정보과학회 학술발표논문집 (2014): 423-425.
[2] U. Shadikhodjaev, et al. "Biaffine Dependency Parser for Korean." Annual Conference on Human and Language Technology. Human and Language Technology, 2018.
[3] 윤준영, et al. “BERT를 이용한 한국어 단어 의미 모호성 해소.”, 한글 및 한국어 정보처리 학술대회, 2019.10.
[4] 이재성. “한국어 형태소 분석을 위한 3 단계 확률 모델.” 정보과학회논문지: 소프트웨어 및 응용 38.5 (2011): 257-268.
[5] 박천음, et al. “BERT 기반 LSTM 모델을 이용한 한국어 형태소 분석 및 품사 태깅.” 한글 및 한국어 정보처리 학술대회, 2019.10.
[6] 이창기. “Structural SVM 을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델.” 정보과학회논문지: 소프트웨어 및 응용 40.12 (2013): 826-832.
[7] 나승훈, et al. “구기반 통계적 모델을 이용한 한국어 형태소 분할 및 품사 태깅.” 한국정보과학회 학술발표논문 집, 2014, pp. 571-573.
[8] 심광섭. “음절 단위의 한국어 품사 태깅에서 원형 복원.” 정보과학회논문지: 소프트웨어 및 응용 40.3 (2013): 182-189.
[9] 이충희, et al. “기분석사전과 기계학습 방법을 결합한 음절 단위 한국어 품사 태깅.” 정보과학회논문지 43.3 (2016): 362-369.
[10] I. Sutskever, et al. “Sequence to sequence learning with neural networks.” Advances in neural information processing systems. 2014.
[11] 이건일, et al. “Sequence-to-sequence 기반 한국어 형태소 분석 및 품사 태깅.” 정보과학회논문지 44.1 (2017): 57-62.
[12] 박건우, et al. “Sequence-to-Sequence 기반 다중 발화 후보를 이용한 형태소 분석기.” 한국정보과학회 학술발표논문집 (2017): 648-650.
[13] J. Devlin, et al. “Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 2018.
[14] A. Vaswani, et al. “Attention is all you need.” Advances in neural information processing systems. 2017.
[15] 국립국어원, “21세기 세종프로젝트 최종성과물, 수정판”, 2011.
[16] 민진우, et al. “BERT에 기반한 Subword 단위 한국어 형태소 분석.” 한글 및 한국어 정보처리 학술대회, 2019.10.