End-to-End 뉴럴 전이 기반 한국어 형태소 분석

End-to-End Neural Transition-based Morpheme Segmentation and POS Tagging of Korean

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

출처 (Source)
한국정보과학회 학술발표논문집 , 2019.6, 566-568(3 pages)

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

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08763255

APA Style
민진우, 나승훈 , 신종훈, 김영길 (2019). End-to-End 뉴럴 전이 기반 한국어 형태소 분석. 한국정보과학회 학술발표논문집, 566-568

End-to-End 뉴럴 전이 기반 한국어 형태소 분석

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

End-to-End Neural Transition-based Morpheme Segmentation and POS Tagging of Korean

Jinwoo Min O1, Seung-Hoon Na2, Jong-Hoon Shin3, Young-Kil Kim4
[12]Chonbuk National University, [34]ETRI

요 약

음절 단위 한국어 형태소 분석은 음절 단위로 복합 형태소 분석을 수행한 후 복합 형태소를 단위 형태소로 분해하는 후처리 과정이 필요하다. 본 논문에서는 전이 기반 방식을 사용하여 형태소 분석을 수행함과 동시에 해당 형태소가 복합 형태소일 경우 단위 형태소로 분석하는 모델을 제안한다. 실험 결과, 세종 형태소 부착 말뭉치 셋에서 형태소 F1 : 97.86%, 어절 정확도 : 96.34%로 기존의 기분석 사전을 이용한 방법보다 각각 F1: 0.13%, 어절 정확도 : 0.17%의 성능 향상을 보였다.

1. 서론

형태소 분석은 입력된 문장 내의 어절들을 뜻을 지니는 최소의 단위인 형태소들로 분리하고 품사 태그를 부착하는 작업이다[1]. 형태소 분석의 부정확한 분석 결과는 구문 분석, 의미역 결정, 질의·응답 등에 치명적인 영향을 미칠 수 있어 올바른 형태소 분석이 매우 중요하다[1]. 기존의 형태소 분석 방법은 음절 단위 형태소 분석 방법[1-6]이 주를 이루었는데 이러한 방식은 원형 복원의 후처리 단계가 필요하고 이러한 후처리 방법으로 학습 데이터에 나타난 기분석 결과를 사전으로 활용하는 방법이 주로 사용되었다[2].

본 연구에서는 전이 기반 방식으로 형태소 분석을 수행하는 전이 기반 한국어 형태소 분석 모델[3]에 기분석 사전[1-2,5]을 이용한 후처리 과정을 필요로 하지 않는 복합 형태소에서 단위 형태소로의 원형 복원을 일괄적으로 학습하는 End-to-End 형태소 분석 모델을 제안한다. 실험 결과, 세종 형태소 부착 말뭉치 셋에서 형태소 F1 : 97.86%, 어절 정확도 : 96.34%로 기존의 기분석 사전을 이용한 방법보다 각각 F1: 0.13%, 어절 정확도 : 0.17%의 성능 향상을 보였다.

2. 관련연구

음절 기반 한국어 형태소 분석은 주로 순차 태깅 기반으로 연구가 진행 되었는데 [4,5]는 각각 기존의 기계학습 모델인 CRF, Structural SVM을 적용하였고 [1]에서는 각각 품사 태깅, 개체명 인식 등 순차 태깅 문제에서 최고의 성능을 보이고 있는 딥러닝 모델인 Bi-LSTM-CRF를 적용하였다. [3]에서는 의존 파싱 등에서 널리 활용되는 전이 기반 방식[7]을 한국어 형태소 분석에 알맞게 확장하여 분석 성능을 높였다.

위의 음절 기반 순차 태깅 모델과 전이 기반 모델은 모두 복합 형태소 단위로 태깅하는 방법을 사용하는데 이는 복합 형태소를 단위 형태소로 분할하여 원형을 복원하는 과정이 필요하다. 복합 형태소를 단위 형태소로 분할하는 가장 기본적인 방법은 학습 데이터에 나타나는 복합 형태소의 패턴들을 사전화하여 복합 형태소를 분해하는 기분석 사전을 이용하는 것이다[1-2,5]. 이러한 기분석 사전 방법은 가변형이 많은 활용형의 용언류에 대해서 발생하는 미등록어 문제에 취약한 단점이 있다. [4, 6]에서는 Lattice HMM 기반 방식을 제안하였는데 위의 방법은 높은 시간 복잡도를 지니며, 형태소 분석 과정을 한 가지 모델로 통일하지 못한다는 단점이 존재한다. [6]에서는 이를 보완하기 위하여 복합 형태소에서 기능 형태소만을 분석하는 모델을 제안한 후 실험 결과를 보인다.

3. 모델

3.1 형태소 분석 전이 액션

본 논문에서 형태소 분석을 위한 전이 액션은 두 가지 액션이고 역할은 다음과 같다.

$Seperate$ Action : 형태소의 끝 경계를 결정하고 해당 형태소의 품사를 결정하는 액션. 버퍼의 Top에 있는 음절을 현재 스택에 Push한 후 품사를 결정한다.

• $shift\ Action$: 현재 음절을 형태소의 요소로 추가하는 액션. 단순히 Top에 있는 음절을 스택에 삽입한다.

표 1. 전이 액션 별 버퍼 및 스택 정보의 갱신 과정

S_t	B_t	Action	S_{t+1}	B_{t+1}
S	c, B	$Seperate(t)$	$(t, c), S$	B
S	c, B	$Shift$	c, S	B

위의 표 1는 형태소 분석의 전이 액션 별 버퍼 및 스택 정보의 갱신 과정을 보여준다. 전이 액션을 위한 버퍼와 스택은 B, S 로 표기하고 기호 c, t 는 각각 음절과 품사 태그로 정의한다. $Seperate$ 액션이 수행되면 버퍼에 있던 음절 c가 버퍼로 삽입되고 음절 품사태그 t 가 부여됨을 알 수 있다. 다음으로 $Shift$ 액션은 형태소에 해당 음절을 추가하는 액션으로 현재 버퍼의 Top의 음절을 스택으로 이동시키는 동작만을 수행한다.

3.2 전이 기반 형태소 분석 모델

버퍼의 입력 표상은 입력열 $\boldsymbol{x} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ 로부터 LSTM을 통해 얻어지게 된다.

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

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

입력 벡터 $\boldsymbol{x}_t$ 는 위의 수식 (1)과 같이 입력 문장의 t 번째 음절 임베딩 벡터를 c_t 라 하고 해당 음절이 어절의 시작인지 아닌지를 나타내는 지에 대한 [B, I]에 대한 태그를 임베딩 벡터로 취한 s_t 의 결합으로 이루어진다. 식(2)에서와 같이 입력 열 $\boldsymbol{x}$을 여러 층의 LSTM을 통해 얻어낸 은닉 열 $\boldsymbol{h} = \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\}$을 버퍼의 입력으로 사용하게 된다.

[3] 에서의 동일한 모델로 전이 액션을 결정하게 되는데 버퍼 B 와 스택 S 그리고 전이 액션을 통해 예측된 형태소의 태그를 저장하기 위한 또 다른 스택 P라 하자.

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

여기서 버퍼 B와 그리고 두 스택 S, P의 Top 노드 상태 표상을 $\boldsymbol{B}_t, \boldsymbol{S}_t, \boldsymbol{P}_t$ 로 정의 하며 위의 수식과 같이 세 상태 표상을 결합한 후 비선형 변환을 통해 태거 상태 표상 $\boldsymbol{T}_t$을 얻는다. 버퍼 B는 Top 2개, 두 스택 S, P는 Top 4개를 취하여 노드 상태표상을 얻으며 P의 각 노드 표상 역시 임베딩을 통해 얻어지게 된다. 얻어진 태거 상태 표상 $\boldsymbol{T}_t$는 소프트맥스 층으로 연결되어 얻어진 확률 값 중 최대가 되는 액션으로 다음 전이를 수행하게 되고 버퍼가 완전히 비워지게 되는 최종 상태에 도달할 때까지 반복하면서 형태소 분석을 수행한다.

3.3 End-to-End 단위 형태소 분석

위의 모델은 복합 형태소 단위로 수행되며 이들 연구에서는 복합 형태소를 처리하기 위한 별도의 과정이 필요하다. 복합 형태소를 처리하기 위한 방법으로 가장 빈번하게 사용되는 방법은 학습 데이터에 나타나는 기분석 패턴을 사전화하여 이를 이용하는 방법이다. [1-2,4-6].

본 연구에서는 Sequence-to-Sequence[8] 모델을 활용하여 복합 형태소를 분해하는 End-to-End 모델로 전이 기반 형태소 분석 모델을 확장한다.

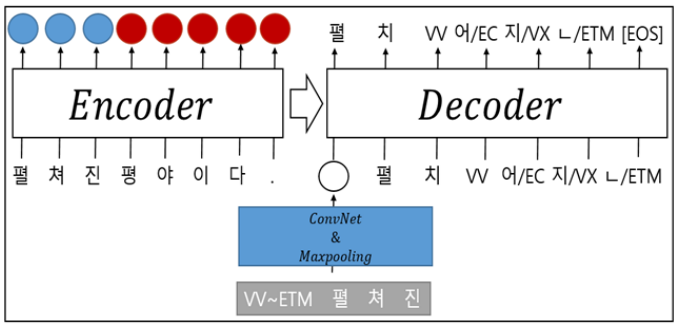

그림 1. Neural Decoder 단위 형태소 생성

복합 형태소를 해결하기 위한 인코더는 3.2절의 식(2)의 은닉열 $\boldsymbol{h} = \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\}$ 을 사용하며 단위 형태소 생성을 위한 디코더를 다음 그림 1과 같이 구성하였다. 먼저 전이 기반 시스템에서 $Seperate$ 액션으로 형태소 "펼쳐진"의 경계가 결정지어짐과 동시에 복합 품사 태그 VV~ETM가 부착될 때 학습 과정에서 해당 형태소가 복합 형태소임을 인식하게 된다. 인식된 복합 형태소는 맨 앞의 내용(content) 형태소와 나머지 복합 기능(compound function) 형태소로 나누어질 수 있고[6] 형태소의 특성에 따라 디코더에서 생성하는 단위를 다르게 하여 디코딩을 수행한다.

그림 2에서 보듯이 먼저 내용 형태소 "펼치/VV"는 용언류의 활용형으로 대부분 미등록어 문제에 취약하기 때문에 이 문제를 완화하기 위해 형태소를 구성하는 음절 단위로 디코딩을 수행한 후 해당 품사를 디코딩한다. 기능 형태소들의 시퀀스인 복합 기능 형태소 [어/EC, 지/VX, ㄴ/ETM]는 미등록어 문제가 거의 발생하지 않아 "형태소/품사태그"의 결합 단위로 디코딩 한다. 기계번역에서 [SOS] (Start of Sentence)를 초기의 입력으로 하는 것과 달리 본 연구에서는 복합 형태소 품사와 형태소의 음절들을 ConvNet을 거친 후 MaxPooling을 취해 얻어진 벡터들을 초기의 입력으로 사용하며 모든 단위 형태소와 [EOS]가 생성될 때까지 디코딩을 수행한다.

또한, 여기서도 어텐션 메커니즘[8]을 사용하여 특정 디코딩 시점에 인코더의 해당 음절들을 집중하도록 적용하였으며 형태소에 해당하지 않는 부분(그림에서의 붉은 색)은 가중치가 반영되지 않도록 하고 "펼, 쳐, 진" (그림에서의 파란색)에 가중치가 집중되도록 설정하였다.

4. 실험

4.1 실험 세팅

본 논문에서 제안한 모델을 평가하기 위해 [4]에서와 동일한 집합인 세종 품사 부착 말뭉치를 사용하였다. 위의 학습 데이터의 202,508 문장 중에서 5000문장을 개발 셋으로 나누어 사용하였다. 품사 태그는 세종 품사 태그[9]를 사용하여 총 42개의 품사태그로 구성되어 있다. 본 논문에서는 복합 형태소 단위로 품사 태깅을 수행하며 총 복합 형태소의 개수는 98개이다.

4.2 실험 결과

본 연구에서 제안한 전이 기반 모델과의 성능 비교를 위한 베이스 라인 모델로 기계 학습 모델인 CRF, Phrase-Based CRF, 그리고 딥러닝 모델인 Bi-LSTM-CRF 모델[1]을 음절 기반 방식으로 한국어 형태소 분석에 적용하였다. 아래 표 2는 실험 결과를 보여준다.

표 2. 전이 기반 형태소 분석 실험 결과

	형태소 F1	어절 정확도
CRF[4]	97.61%	96.14%
CRF(revised)	96.63%	96.18%
Phrase-Based CRF [10]	97.74%	96.35%
Bi-LSTM-CRF	96.96%	N/A
전이기반	**97.96%**	**96.72%**

표 2에서 보듯이 전이 기반 형태소 분석 모델이 베이스 라인 모델 중 가장 높은 성능을 보이고 있는 Phrase-Based CRF에 비해 F1 기준 0.15%, 어절 정확도에서 0.31% 높은 성능을 보이고 있음을 알 수 있다.

다음으로는 제안 End-to-End 전이 기반 형태소 분석 모델의 실험 결과를 표 3에 제시한다.

표 3. 단위 형태소 분석 실험 결과

	형태소 F1	어절 정확도
CRF+기분석 [4]	97.08%	95.06%
CRF+기분석(+lattice HMM) [4]	97.21%	95.22%
전이기반+기분석 사전	97.55%	96.17%
전이기반+단위 형태소 생성	**97.68%**	**96.34%**

제안 모델에 대한 베이스 라인으로 전이 기반 모델의 복합 형태소 출력 결과에 학습 데이터에 나타난 복합 형태소들의 기분석 패턴을 사전화한 후 이를 이용하여 단위형태소로 분해한 방식을 사용한다. 사전을 구성할 때 복합 형태소가 여러 단위 형태소들의 조합으로 나타날 수 있는데 가장 빈도수가 높은 단위 형태소들을 선택하도록 하였다. 추가로 [4]의 CRF 형태소 분석 모델에서 기분석, 기분석+lattice HMM 모델을 사용한 복합 형태소 후처리 결과 성능을 제시한다. 평가 지표는 동일하게 단위 형태소 F1, 어절 정확도이다.

표 3에서 보듯이 전이 기반 End-to-End 단위 형태소 생성 모델의 결과가 출력된 복합 형태소를 기분석 사전을 통해 분해한 결과보다 형태소 F1 0.13%, 어절 정확도 0.17% 높은 결과를 보임을 알 수 있다.

5. 결론

본 연구에서는 전이 기반 방식을 한국어 형태소 분석 테스크에 적용하여 기존의 형태소 분석 성능을 뛰어넘는 결과를 보여주었고 기분석 사전 방식이 아닌 Sequence-to-Sequence 모델을 적용하여 단위 형태소 분할까지 딥러닝 모델 상에서 일괄적으로 처리하는 End-to-End 형태소 분석 모델을 제안하여 기분석 사전을 이용한 후처리 결과보다 높은 성능을 보일 수 있음을 보였다.

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] 신준철, 옥철영. (2012). 기분석 부분 어절 사전을 활용한 한국어 형태소 분석기. 정보과학회논문지 : 소프트웨어 및 응용, 39(5), 415-424.

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

[4] Na, S. H. (2015). Conditional random fields for korean morpheme segmentation and pos tagging. ACM Transactions on Asian and Low-Resource Language Information Processing, 14(3), 10.

[5] 이창기. "Structural SVM 을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델." 정보과학회논문지: 소프트웨어 및 응용 40.12 (2013): 826-832.

[6] 나승훈, 김창현, 김영길. “CRF 기반 한국어 형태소 분할 및 품사 태깅에서 두 단계 복합형태소 분해 방법” HCLT, 2013.

[7] Dyer, C., Ballesteros, M., Ling, W., Matthews, A., & Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. arXiv preprint arXiv:1505.08075.

[8] 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).

[9] 홍진표, & 차정원. (2013). 품사 태거와 빈도 정보를 활용한 세종 형태 분석 말뭉치 오류 수정. 정보과학회논문지: 소프트웨어 및 응용, 40(7), 417-428.

[10] 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)