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, kimyk}@etri.re.kr
Cognitive Computing Lab.

Ⅰ. 서론

형태소 분석 문장 내의 어절들을 뜻을 지니는 최소의 단위인 형태소들로 분리하고 품사태그를 부착하는 작업.

・음절 단위 형태소 분석

음절 단위 형태소 분석은 원형 복원의 후처리 단계가 필요하고 이러한 후처리 방법으로 학습 데이터에 나타난 복합 형태소의 기분석 결과를 사전으로 활용하는 방법이 주로 사용됨.

거리는	→	거리 [NNG] 는 [JX]
사람의	→	사람 [NNG] 의 [JKG]
물결로	→	물결 [NNG] 로 [JKB]
넘쳤다.	→	넘쳤 [VV~EP] 다 [EF] . [SF]

Ⅱ. 제안 방법

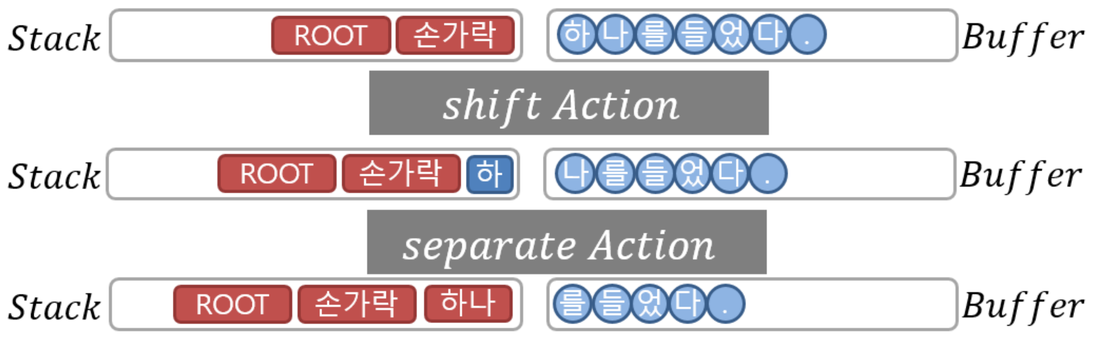
Proposed System

본 연구에서는 전이 기반 방법을 사용하여 음절 단위의 복합 형태소 분석을 수행 한 후 인식된 복합 형태소를 Sequence-to-Sequence 모델을 이용하여 단위 형태소로 분리하는 End-to-End 형태소 분석 모델을 제안.

ⅰ.형태소 분석 전이 액션

・Shift Action 현재 음절을 형태소의 요소로 추가하는 액션. 단순히 Top에 있는 음절을 스택에 삽입.

・Seperate Action 현재 형태소의 끝 경계를 결정하고 해당 형태소의 품사를 결정하는 액션. 버퍼의 Top에 있는 음절을 현태 스택에 Push한 후 품사를 결정.



ⅱ.전이 기반 형태소 분석 모델

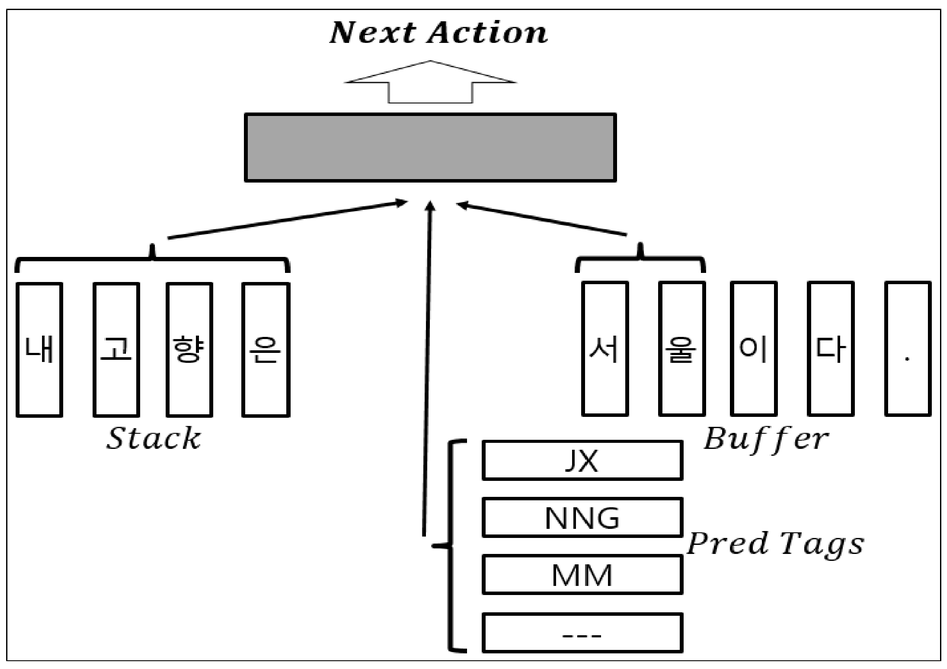
버퍼의 입력 표상은 음절에 대한 입력열 $\boldsymbol{x} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$로부터 LSTM을 통해 인코딩 하여 얻어지고 입력 벡터 $\boldsymbol{x}_t$는 음절과 해당 음절이 어절의 시작인지 아닌지에 대한 [B, I] 태그로 구성.

$$\boldsymbol{x}_t = [\boldsymbol{c}_t; \boldsymbol{s}_t]$$

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

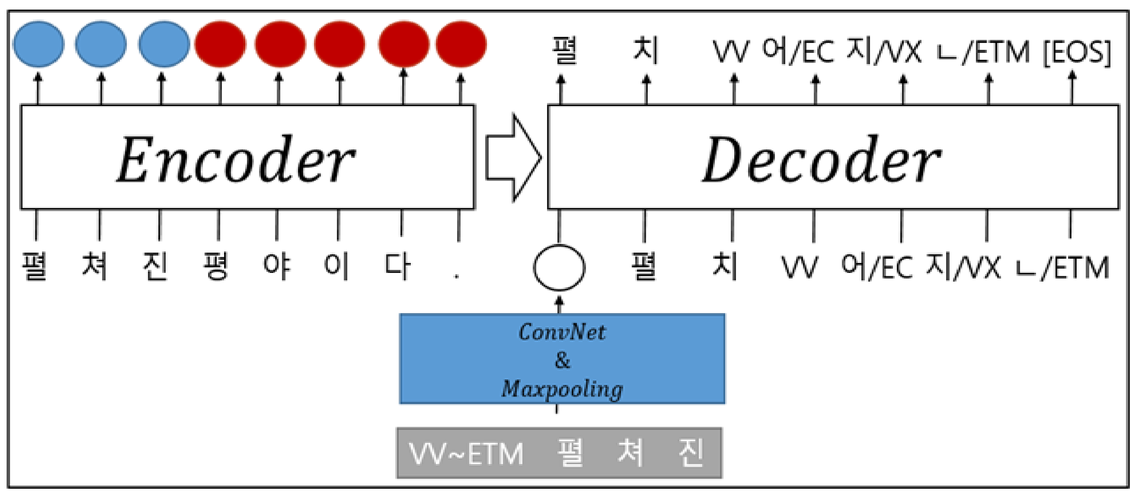
현재 버퍼와 스택 그리고 예측된 태그를 저장하기 위한 스택으로부터 자질을 추출한 후 다음 전이 액션을 결정.

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



ⅲ. End-to-End 단위 형태소 생성

전이 기반 형태소 분석 모델로 인식된 형태소가 복합 형태소인 경우 이를 Sequence-to-Sequence 모델을 이용하여 단위 형태소로 분석하는 End-to-End 모델로의 확장.



・복합형태소 분석 복합 형태소는 맨 앞의 내용(content) 형태소와 나머지 복합 기능 (compound fucntion) 형태소로 나뉠 수 있음. 내용 형태소는 용언 류의 활용형으로 미등록어 문제를 해결하기 위해서 음절단위로 디코딩 한 후 품사 태그를 디코딩. 복합 기능 형태소는 미등록어 문제가 거의 발생하지 않아 "형태소/품사태그"의 결합 단위로 디코딩 수행.

・초기 입력 기계번역에서 SOS(Start of Sentence)를 초기 입력으로 하는 것과 달리 복합 형태소의 품사와 형태소의 음절들을 ConvNet을 거친 뒤 MaxPooling 하여 얻어진 벡터를 사용.

・Masked Attention 어텐션 메커니즘을 이용하여 각 디코딩 시점에 인코더 음절에 집중할 때 분리하고자 하는 형태소에 해당하지 않는 부분에 가중치가 반영되지 않도록 Mask처리 하여 어텐션 스코어가 형태소에 해당하는 부분만 집중하도록 처리.

Ⅲ. 실험 결과

・실험 집합 세종 품사 부착 말뭉치를 사용하며 학습 데이터의 202,508 문장 중에서 5000문장을 개발 셋으로 나누어 사용. 품사 태그는 세종 태그를 사용하며 총 42개의 품사태그로 구성

・전이 기반 형태소 분석 복합 형태소 단위로 형태소 분석을 수행하며 베이스라인으로 CRF, Phrase-Based CRF, 딥러닝 모델인 Bi-LSTM-CRF 모델을 제시

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

・단위 형태소 분석 전이 기반 End-to-End 단위 형태소 생성 모델이 기존의 기분석 사전을 이용한 방법보다 형태소 F1 : 0.13%, 어절 정확도 : 0.17% 높은 성능을 보임

모델	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%**

Ⅳ. 결론

본 연구에서는 전이 기반 방법을 사용하여 음절 단위의 복합 형태소 분석을 수행 한 후 인식된 복합 형태소를 Sequence-to-Sequence 모델을 이용하여 단위 형태소로 분리하는 End-to-End 형태소 분석 모델을 제안하고 기존의 기분석 사전을 이용한 방법보다 높은 성능을 보임