인터리브-디코더를 이용한 Sequence-to-Sequence기반 한국어 형태소 분석

Cognitive Computing Lab.

민진우[O1], 나승훈[2], 신종훈[3], 김영길[4]
[12] 전북대학교, [34] 한국전자통신연구원
jinwoomin4488@gmail.com, nash@jbnu.ac.kr, {jhshin82, kimyk}@etri.re.kr

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.

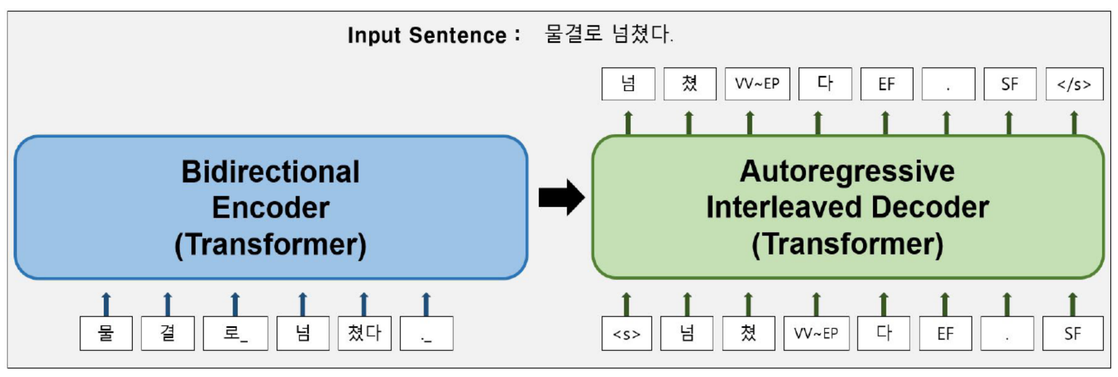
거리는	→	거리 [NNG] 는 [JX]
사람의	→	사람 [NNG] 의 [JKG]
물결로	→	물결 [NNG] 로 [JKB]
넘쳤다.	→	넘쳤 [VV~EP] 다 [EF] . [SF]

Ⅱ. 제안 방법

Proposed System

본 연구에서는 BERT 등과 같은 사전 훈련된 인코더의 파라미터를 디코더에서 효율적으로 초기화하기 위해 기존의 트랜스포머 구조를 인코더와 일관성 있는 구조로 변형 인터리브-디코더를 이용한 형태소 분석 시스템 제안.

ⅰ. 한국어 형태소 분석을 위한 시퀀스-투-시퀀스 모델과 입출력



▪ 인코더로는 양방향 트랜스포머를 사용하고 한국어 ETRI BERT를 통해 초기화. 디코더는 기존의 디코더를 변형한 인터리브 디코더를 사용하고 역시 인코더와 동일하게 ETRI BERT의 인코더 파라미터를 활용하여 초기화

▪ BERT 모델을 활용하기 위해 음절 시퀀스가 아닌 한국어 BERT의 입력인 워드 피스 단위로 분리한 토큰들을 입력으로 하고 어절의 마지막 토큰임을 알리는 "_"가 부착된 형태

▪ BERT 모델을 활용하기 위해 음절 시퀀스가 아닌 한국어 BERT의 입력인 워드 피스 단위로 분리한 토큰들을 입력으로 하고 어절의 마지막 토큰임을 알리는 "_"가 부착된 형태

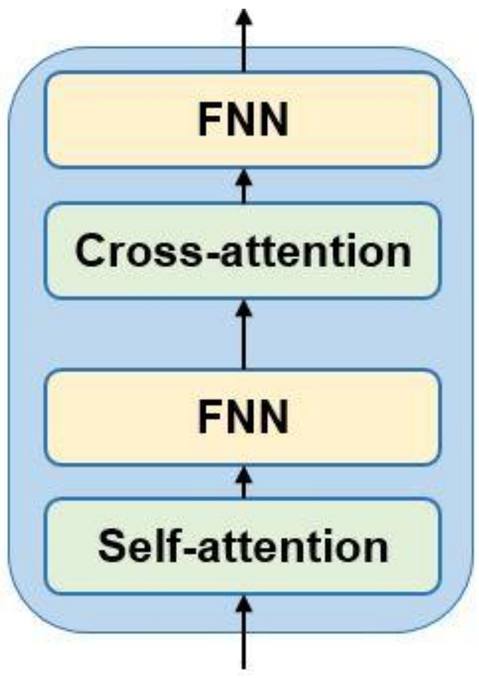
▪ 음절 단위로 형태소 분석을 수행 : 기존 연구들과의 성능 비교를 위함. 음절 단위로 형태소 분석을 수행 출력층에서 음절과 복합 품사의 시퀀스를 출력. 여기서 품사가 나타날때까지 연속된 음절의 모음이 하나의 형태소 가 되며 뒤에 나타난 품사가 얻어진 형태소의 품사

ⅱ. 인터리브 디코더

인터리브-디코더는 BERT 등과 같은 사전 훈련된 인코더의 파라미터를 디코더에서 효율적으로 초기화하기 위해 기존의 트랜스포머 구조를 인코더와 일관성 있는 구조로 변형.

▪ 기존의 트랜스포머 디코더는 [셀프 어텐션 층, 크로스 어텐션 층, FNN(Feedforward Neural Networks) 층] 순으로 서브 층를 구성

▪ **제안 인터리브-디코더** 사전 훈련된 인코더와 일관성을 유지하도록 셀프 어텐션 레이어 이후에 FNN층을 추가하여 각 트랜스포머 블록은 하나의 셀프 어텐센, 하나의 크로스 어텐션 및 두 개의 FFN으로 구성



▪ **파라미터 초기화** 사전학습 인코더의 홀수 층의 파라미터로 셀프 어텐션 층과 하위 FNN층을 초기화하고 짝수 층의 파라미터로 크로스 어텐션 층과 상위 FNN층을 초기화. ETRI BERT는 12층으로 구성된 BERT-base와 동일한 세팅. 본 논문에서는 인코더의 두개의 층으로 하나의 디코더 파라미터를 초기화하기 때문에 디코더 층은 6개로 설정

Ⅲ. 실험 결과

▪ **실험 집합 & 평가 지표** 세종 형태소 분석 말뭉치를 이용. 학습 셋 202,508 문장, 평가 셋 52,781 문장으로 구성. 학습 셋에서 5,000문장을 별도로 나누어 개발셋으로 사용. 평가지표로는 형태소 단위 F1 점수와 어절 내의 모든 형태소가 올바르게 분석되었는지에 대한 어절 정확도를 제시

▪ **4가지의 초기화 세팅 실험** 인터리브-디코더 구조와 초기화 단계에 따른 성능 변화를 측정. "Load"은 기학습 파라미터를 로딩함을 의미하고 "Random"은 기학습 파라미터를 로딩하지 않고 파라미터를 랜덤 초기화함을 의미

1) 인코더 [Random] – 디코더 [Random]
2) 인코더 [Load] - 디코더 [Random]
3) 인코더 [Load] - 디코더 [Load]
4) 인코더 [Load] - 인터리브 디코더 [Load]

▪ **4가지의 초기화 세팅 실험** 인터리브-디코더 구조와 초기화 단계에 따른 성능 변화를 측정. "Load"은 기학습 파라미터를 로딩함을 의미하고 "Random"은 기학습 파라미터를 로딩하지 않고 파라미터를 랜덤 초기화함을 의미

모델	형태소 F1	어절 정확도
(BERT) + 전이기반	98.01%	96.78%
(BERT) + Bi-LSTM CRF	98.03%	96.81%
스택-포인터 네트워크	98.12%	96.92%
인코더[Random]-디코더[Random]	97.12%	95.60%
인코더[Load]-디코더[Random]	98.14%	97.02%
인코더[Load]-디코더[Load]	98.18%	97.07%
인코더[Load]-인터리브 디코더[Load]	**98.19%**	**97.07%**

▪ **결과 분석** 인코더 파라미터를 활용하여 디코더를 초기화하였을 때의 효과는 형태소 F1 : 0.04%p, 어절 정확도 : 0.05%p로의 성능 향상. 기존의 디코더와 인터리브-디코더에서의 성능은 거의 차이가 존재하지 않아 다양한 초기화 방법에 대한 실험을 진행할 예정

Ⅳ. 결론

본 논문에서는 생성 기반 한국어 형태소 분석 모델에서 사전 학습 언어모델의 파라미터를 디코더에서 인코더와 일관성을 유지하도록 디코더 구조를 변형한 인터리브 디코더를 이용한 형태소 분석 실험 진행. 향후 연구로는 형태소의 원형 문자들과 단위 형태소를 End-to-End로 생성하는 실험을 진행할 예정.