# End-to-End 뉴럴 전이 기반 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

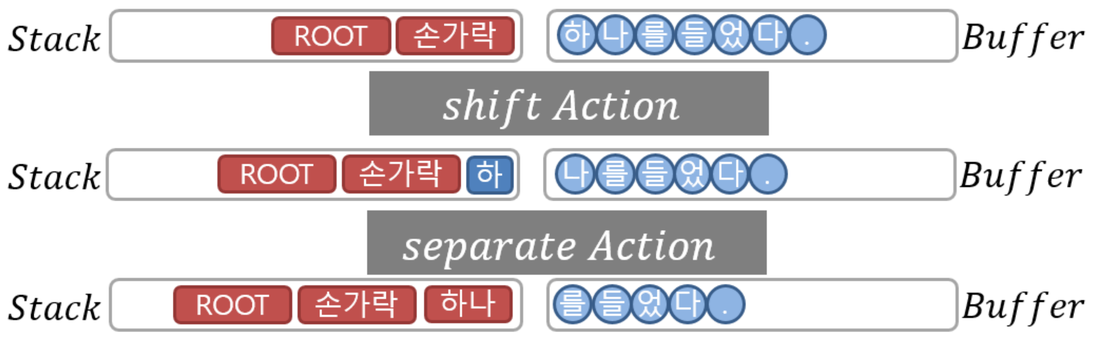
### Proposed System

본 연구에서는 전이 기반 방법을 사용하여 음절 단위의 복합 형태소 분석을 수행하는 전이 기반 형태소 분석 모델에 최근 다양한 자연어 처리 분야에서 높은 성능 향상을 보이고 있는 BERT 모델을 한국어 형태소 분석 태스크에 적용하여 실험결과를 보임

**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.

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

| 입력 문장 |
|---|
| 나는 오늘 똑똑히 보았단다. |
| 음절 단위 |
| 나, 는, 오, 늘, 똑, 똑, 히, 보, 았, 단, 다, . |
| KorBERT 모델의 입력 |
| [CLS] 나, 는_, 오, 늘_, 똑, 똑, 히_, 보, 았, 단, 다, ._, [SEP] |

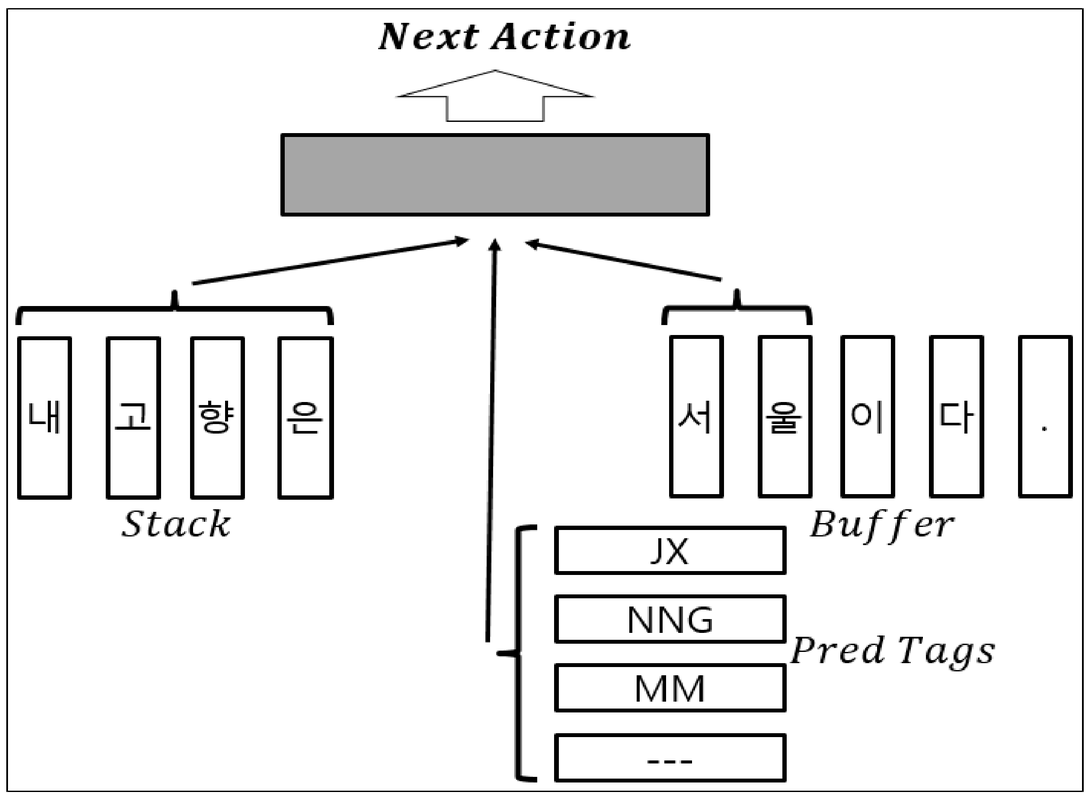
**ⅲ. 전이 기반 형태소 분석 모델**

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 음절 임베딩과 해당 음절이 어절의 시작인지 아닌지에 대한 [B, I] 태그 임베딩 $\boldsymbol{s}_t$ 이외에 본 연구에서 제안한 BERT 모델을 통해 인코딩된 $\boldsymbol{b}_t$를 추가적으로 사용

$$\boldsymbol{x}_t = [\boldsymbol{c}_t; \boldsymbol{s}_t; \boldsymbol{b}_t]$$
$$\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\} = LSTM(\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\})$$

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. 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

· **베이스 라인 모델** 베이스 라인으로 전이 기반 모델을 포함한 기존 연구 이외에 Bi-LSTM-CRF 모델과 이에 BERT를 적용한 두 모델을 제시 이는 은닉 표상 $\boldsymbol{h}_i$에 대해 MLP를 적용 한 후 출력 층에서 태그 간의 의존성을 모델링하는 CRF와 결합한 모델

· **실험 결과**

| 모델 | 형태소 F1 | 어절 정확도 |
|---|---|---|
| CRF | 97.60% | 96.14% |
| Phrase-Based-CRF | 97.74% | 96.35% |
| 전이기반 | 97.91% | 96.65% |
| subword BERT + LSTM | 95.22% | 93.90% |
| 전이기반(re-impl) | 98.01% | 96.78% |
| Bi-LSTM-CRF(impl) | 98.03% | 96.81% |
| 전이기반(re-impl) + BERT | 98.01% | 96.72% |
| Bi-LSTM-CRF(impl) + BERT | 98.01% | 96.75% |

## Ⅳ. 결론 및 향후 연구

본 연구에서는 전이 기반 모델에 음절 단위 BERT를 적용하여 실험 결과를 얻었고 전반적으로 성능에 큰 변화가 없었고 어절 단위 KorBERT 모델은 근본적으로 음절 단위 형태소 분석에 적합한 음절 단위로 학습한 BERT가 아니기 때문에 BERT 모델을 적용했을 때의 형태소 분석 성능 향상이 미미함.

음절 단위 BERT를 포함하여 향상된 모델인 XLNet, 후속 모델들에 대해 음절 정보를 효율적으로 학습할 수 있는 모델에서 학습 후 형태소 분석에 적용할 예정