# BERT에 기반한
# Subword 단위 한국어 형태소 분석

민진우[1], 나승훈[1], 신종훈[2], 김영길[2]
[1]전북대학교 인지컴퓨팅연구실, [2]ETRI

# 목차

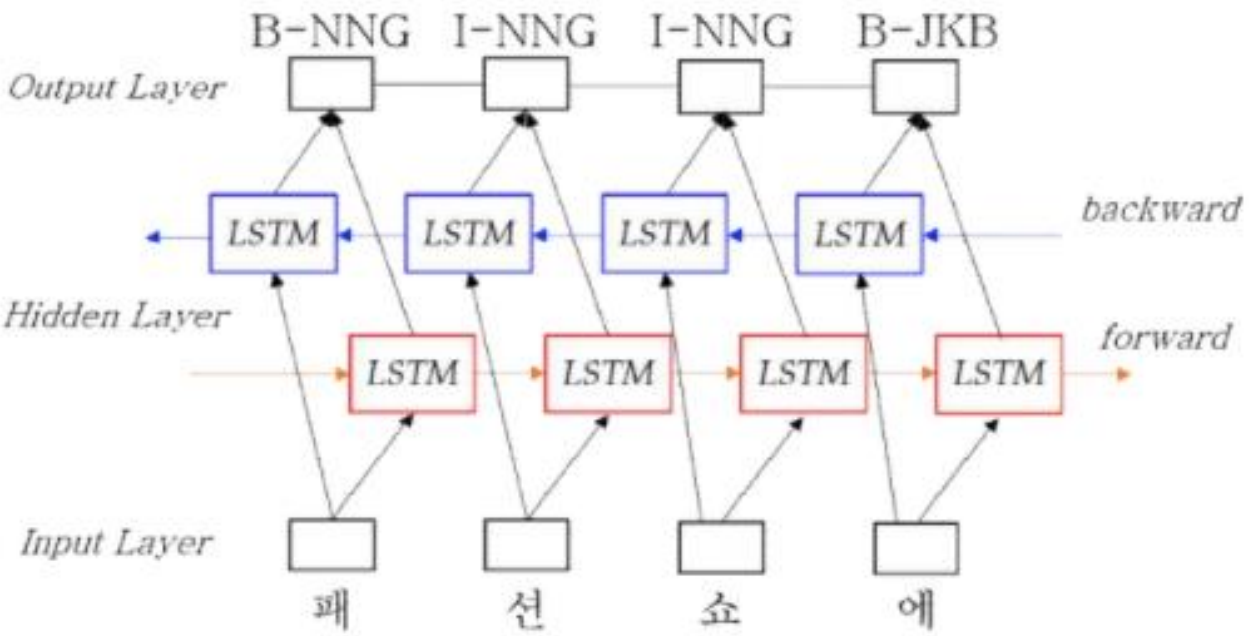
- 형태소 분석
- 관련연구
- BERT에 기반한 Subword 단위 한국어 형태소 분석
- 실험결과

# 형태소 분석

- 정의(두 가지의 과정으로 구분)
  - 형태소 분석 : 문장 내의 어절을 뜻을 지니는 최소의 단위인 형태소로 분해하고 해당 형태소의 품사 후보를 생성
  - 품사 태깅 : 형태소의 품사 후보로부터 가장 적절한 품사를 결정하는 과정
  - 입력문
    - 예) 거리는 사람의 물결로 넘쳤다

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

# 품사 분포와 Bidirectional LSTM-CRFs를 이용한 음절단위 형태소 분석기(김혜민, HCLT '2016)



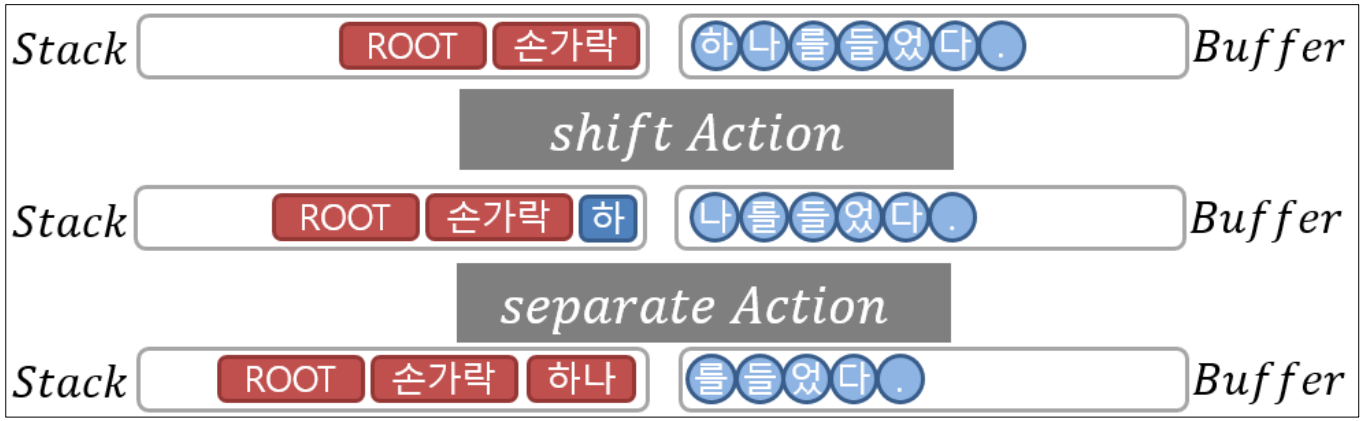

- Bi LSTM CRFs 형태소 분석
  - 음절 단위의 품사 태깅 방법
  - 순차 데이터를 모델링하는 양방향 LSTM에 출력 태그 간의 전이 확률을 얻는 CRF와 결합하는 방식
  - [B(Begin),I(Inside)] 등의 태그 등을 붙인 품사 태그를 결정하는 방식

# 동적 오라클을 이용한 뉴럴 전이기반 한국어 형태소 분석 및 품사 태깅(민진우, HCLT '2018)

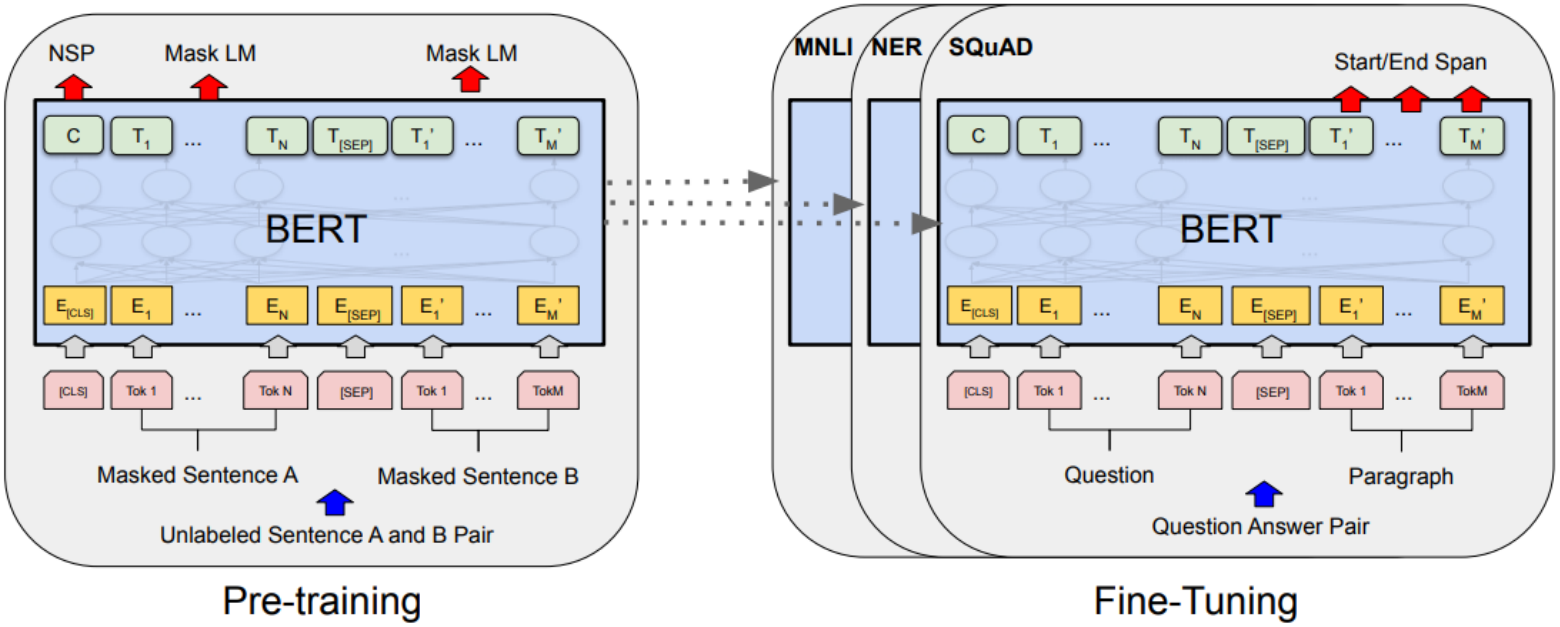
- 전이 기반 형태소 분석 모델
  - 두가지의 Action으로 구성
    - $separate$ : 현재 음절을 형태소의 요소로 추가하는 Action
    - $shift$ : 형태소의 끝 경계를 결정하고 품사를 결정하는 Action

| $S_t$ | $B_t$ | Action | $S_{t+1}$ | $B_{t+1}$ |
|---|---|---|---|---|
| $S$ | $c, B$ | $Separate(t)$ | $(t, c), S$ | $B$ |
| $S$ | $c, B$ | $shift$ | $c, S$ | $B$ |

  - 실행 예




# Bert: Pre-training of deep bidirectional transformers for language understanding(Devlin, J. NAACL '2019)




- BERT : Bidirectional Encoder Representation from Transformer
  - 양방향의 Transformer를 이용하여 문장 내 임의의 단어를 예측하고 다음 문장을 예측하는 두 가지 Task로 언어 모델 학습
- 응용 Task에 Fine-Tuning하는 방식으로 성능 향상

# Bert: Pre-training of deep bidirectional transformers for language understanding(Devlin, J. NAACL '2019)

- BERT-base
  - 영문에 적용하기 위해서 문장을 Space(공백) 단위로 Tokenize하여 학습한 모델
- BERT-multilingual
  - 다양한 언어를 지원하기 위하여 입력을 word piece(sub-word)단위로 Tokenize 한 후 BERT를 학습한 모델

```
tokenizer = BertTokenizer.from_pretrained(
    "bert-base-multilingual-cased", do_lower_case=False)
tokenizer.tokenize("Elvégezhetitek")

['El', '##vé', '##ge', '##zhet', '##ite', '##k']
```

# BERT에 기반한 Subword 단위 한국어 형태소 분석

- 3 단계로 구성

**전처리**
1.학습데이터의 문장을 Sub-word 단위로 분리
2. 음절 단위 태그를 Sub-word 단위의 복합태그로 합성

**모델 적용**
Subword 단위
BERT기반 LSTM 형태소 분석 모델 적용

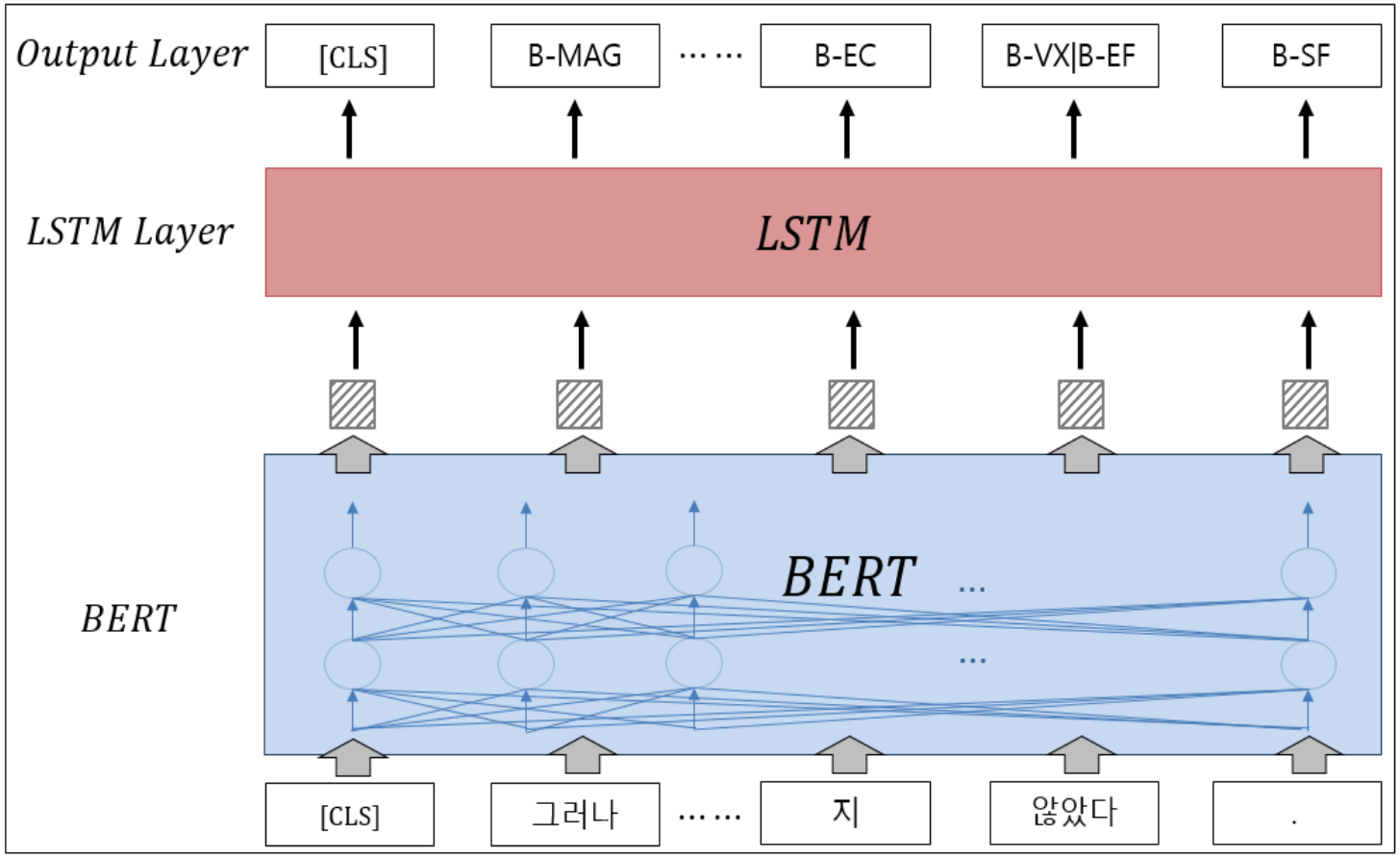
**후처리 및 평가**
1. sub-word 단위 복합 태그 음절 단위로 분해
2.형태소 분석 평가

- Subword 단위의 전처리
  - 입력 문장에 대해 BERT-multiliugual wordpiece 토크나이저를 사용하여 Subword로의 토큰화
  - 대부분이 단일 음절로 구성된 Subword
  - 한국어 입력의 토큰화 예)

| | |
|---|---|
| 문장 | 가로수 잎들이 드높게 펄럭이고 있었다. |
| 토큰화된 문장 | 가 ##로 ##수 잎 ##들이 드높게 펄럭 ##이고 있었다 ##. |
| 문장 | 어쩜 이렇게 통할 수가 있을까. |
| 토큰화된 문장 | 어쩜 이 ##렇게 통 ##할 수 ##가 있을 ##까## . |

- Subword 단위의 전처리
  - 음절 단위 태그를 sub-word 단위의 복합태그로 합성
    - 합성 규칙
      1. 토큰화된 sub-word가 단일 음절이면 해당 음절의 태그를 그대로 사용
      2. 토큰화된 sub-word가 복합 음절로 구성되지만 태그가 같으면 단일 태그 사용
      3. 토큰화된 sub-word가 복합음절이면서 여러 태그로 구성되면 시작 음절의 태그와 끝 음절의 태그의 결합으로 구성
  - 복합 태그의 구성 예

| 예제 | | |
|---|---|---|
| ① | sub-word | 꿈 [B-NNG] |
| | 복합 태그 | B-NNG |
| ② | sub-word | 하지만 [B-MAG,I-MAG,I-MAG] |
| | 복합 태그 | B-MAG |
| ③ | sub-word | 스트 [I-NNG I-NNG] |
| | 복합 태그 | I-NNG |
| ④ | sub-word | 에서는[B-JKB,I-JKB,B-JX] |
| | 복합 태그 | B-JKB\|B-JX |




- sub-word 단위로 토큰화된 문장이 입력 열이 BERT의 입력
- BERT 모델의 출력과 Sub-word가 어절의 시작인지 아닌지를 나타내는 띄어쓰기 태그[B,I]와 결합하여 Bi-LSTM 통해 한번 더 인코딩
- 출력 층에서 sub-word에 대해 확률이 최대가 되는 복합태그를 결정

- sub-word 복합형태소 분해
  - 분석 결과는 sub-word 단위로 음절 단위 태그로 변환
  - 변환 규칙
    1. 토큰화된 sub-word가 단일 음절이면 해당 태그가 해당 음절의 태그로 복원
    2. 토큰화된 sub-word가 2음절로 구성될 때 단일 태그이면 해당 품사를 그대로 부착하고 복합태그이면 복합태그의 시작태그와 끝 태그를 첫 음절과 끝 음절에 각각 부착
    3. 3음절 이상이며 단일 태그가 아닌 경우 기분석 사전에 의한 복원
    4. 3)의 기분석 사전에 존재하지 않는 경우는 휴리스틱하게 복원

- 기분석 사전
  - Key : (음절수, 복합태그)의 튜플
  - Value : (빈도수, 음절태그의 리스트)의 리스트
  - 기분석 사전에 존재하지 않는 4번의 경우 주로 형태소 분석의 오류가 됨

# 실험 세팅

- 데이터 셋
  - 세종 형태소 분석 말뭉치

| | Train | Dev | Test |
|---|---|---|---|
| 문장 수 | 197508 | 5000 | 50631 |
| 어절 수 | 2674563 | 97292 | 694523 |

- 파라미터

| | Hyper Parameter | value |
|---|---|---|
| BERT | 인코더 블록 개수 | 12 |
| | 은닉 차원수 | 768 |
| | 어텐션 헤드 수 | 12 |
| | Optimizer | Adam |
| | 학습률 | $5e^{-5}$ |
| LSTM | RNN 은닉 차원 수 | 512 |
| | RNN Layers | 5 |
| | 학습률 | 0.001 |
| | 드랍아웃 | 0.33 |

- 평가지표
  - 복합 형태소 단위 F1과 어절 정확도를 제시

# 실험 결과

- 형태소 분석 실험 결과

| | 형태소 F1 | 어절 정확도 |
|---|---|---|
| CRF[3] | 97.60% | 96.14% |
| Phrase-Based CRF[4] | 97.74% | 96.35% |
| Bi-LSTM-CRF[12] | 96.96% | N/A |
| 전이기반 모델[12] | 97.91% | 96.65% |
| sub-word Bi-LSTM | 94.38% | 92.71% |
| BERT sub-word Bi-LSTM | **95.22%** | **93.90%** |

- 결과 분석
  - BERT를 사용함의 효과는 명확하나 현재 음절 단위의 전이 기반 모델에 비해 2% 낮은 성능을 보이고 있음
  - 복합 태그가 총 1,100여개로 CRF를 적용함에 있어 시간 복잡도 및 메모리 용량의 한계
  - 현재 Bert-Multilingual 모델에서 한국어 코퍼스는 소규모로 제한되어 있어 대규모 코퍼스에서의 학습 필요

# Q&A

감사합니다.