KorBERT 모델 배포 FAQ (v20190619)

2019-06-19

(1)
[KorBERT 모델] Korean_BERT_Morphology 모델 사용 시, tokenizer 만 변경하면 되나요?
(답변)
Korean_BERT_Morphology 모델은 아래 예제와 같이, 입력 문장에 대해 형태소 분석한 결과를 입력으로 받습니다.

- 원문: ETRI에서 한국어 BERT 언어 모델을 배포하였다.
- 입력 예제: ETRI/SL 에서/JKB 한국어/NNP BERT/SL 언어/NNG 모델/NNG 을/JKO 배포/NNG 하/XSV 었/EP 다/EF ./SF

001_bert_morp_pytorch 폴더의 src_examples 내용을 참고하시면, OpenAPI를 이용한 형태소분석 및 처리 방법을 확인하실 수 있습니다.

(2)
[KorBERT 모델] Korean_BERT_Morphology 모델 사용 시, OpenAPI의 형태소분석 API만 이용해야 하나요?
(답변)
형태소분석기는 TTA 표준 형태소 태그셋(TTAK.KO-11.0010/R1)에 호환되는 형태소분석기 사용이 필요합니다.
예를 들어, 아래 예제와 같은 경우 TTA 가이드라인에서는 전자를 따르고 있습니다.
1) 사용하다: 사용/NNG + 하/XSV 〈-〉 사용하/VV
2) 산다: 살/VV + ㄴ다/EF 〈-〉 산/VV + 다/EF
3) 연구원: 연구/NNG + 원/XSN 〈-〉 연구원/NNG
TTA 표준 가이드라인과 다른 분석 결과를 사용하면 성능에 영향을 미칠 수 있습니다.

(3)
[KorBERT 모델] Korean_BERT_WordPiece 모델을 사용하여, 기계독해(MRC) 태스크에 적용 시 후처리를 적용해야 하나요?
(답변)
WordPiece 모델은 형태소분석을 수행하지 않는 모델로, 조사/어미와 같은 음절이 선행 음절과 결합되는 경우가 자주 발생합니다. (예: 구성된다 → 구 + ##성된 + ##다)
예를 들어, WordPiece 모델에서는 “단어는” 과 같은 어절을 “단”과 “어는” 처럼 형태소와 다른 단위로 구분합니다. 따라서, 기계독해 모델의 정답이 “단어”일 경우, “단”과 “어는”이라는 wordpiece를 정답 경계로 인식 후, 조사 “는”을 필터링하는 단계가 필요합니다.
구체적인 후처리 규칙은 사용하시는 기계독해 데이터의 dev 셋에서, 시스템 결과와 정답 결과를 비교하여 보시면 후처리 대상 규칙을 정리하실 수 있습니다.