실험 결과 보고

언어학과 정대용
2022.02.08

1 개요

컴퓨터공학부 졸업논문으로 작성한 "품사 고려 알고리즘을 통한 CoSDA-ML"의 성능을 분석, 향상시킨다. 구체적으로는 다음 세 가지의 내용으로 구성된다.

(a) 사전 빌드 후 품사고려-CoSDA-ML 의 학습으로 곧바로 이어지는 도구 개발
(b) XLM-R 로의 언어 모델 교체 적용
(c) 여러 언어에서 성능 분석

2 개발 / 실험 결과

a. 도구 개발

언어: 파이썬
사용 라이브러리: NLTK, GoogleTrans
방식: 스크립트 파일 작성

GoogleTrans 라는 파이썬 라이브러리를 활용해 원하는 언어를 입력하면, 영어 훈련 세트에 대해 자동으로 품사 사전과 입력한 언어의 사전을 구축해준다. 옵션에 따라 곧바로 품사고려-CoSDA-ML 로 이어서 학습할 수 있다.

b. XLM-R 로의 교체

기존 CoSDA-ML 에서 import 하는 XLM 및 M-BERT 관련 도구들과, 형상과 같은 파라미터를 XLM-R 에 맞게 재조정하였다.

c. 여러 언어에서 성능 분석

기존 언어는 카탈루냐어(ca), 바스크어(eu)였고, 여기에 다섯 개의 언어를 추가하였다. XLM-R 기준, 학습을 위해 사용된 데이터량이 10G 미만에 해당하여 비교적 데이터가 부족한 언어 중, 감성분류 테스트 셋을 구할 수 있는 경우를 선택 기준으로 하였다. 추가한 언어는, 우즈벡어(uz), 우르두어(ur), 웨일즈어(cy), 스웨덴어(sw), 슬로베니아어(sl)이다.

실험 결과는 다음과 같다.

	mBERT	CoSDA-mBERT	PACoSDA-mBERT	PACoSDA-XLMR
ca	72.56	72.91	74.58	80.34
eu	88.24	90.78	91	92.57
uz	74.02	78.1	78.9	83.5
ur	62.22	64.7	65.8	66.7
cy	72.34	73.5	72.1	82.1
sl	64.1	70.2	68	78.3
sw	75	75.9	77.6	89.2

좌측부터, 바닐라 mBERT / CoSDA-ML / 품사고려 CoSDA-ML / 품사고려 CoSDA-ML – XLMR 을 의미한다. 붉은색은 Baseline 인 CoSDA-ML 에 비해 성능이 감소한 경우, 옅은 연두색은 증가하였으나 큰 폭으로 증가하지는 않은 경우, 초록색은 큰 폭으로 증가한 경우를 의미한다.

이를 통해 품사고려 CoSDA-ML 의 성능이 여러 언어에서 기존 CoSDA-ML 의 성능을 개선하였음을 알 수 있으며, 특히 강력한 언어 모델인 XLM-R 을 통해 매우 큰 폭으로 성능을 더욱 향상시킬 수 있었음을 확인할 수 있다.