졸업작품 프로젝트

아이디어 제안서

김지은 | 채민형 | 최병규 | 최지원

목차
INDEX

01 작품명 및 설명

01 작품명

딥러닝 기반 한국어 문장/문단의 토픽 유추

02 작품 설명

01

모델링

개체명 인식을 위한
딥러닝 모델을 구현한다.
형태소 분석기를 사용해
개체명 인식이
가능하도록 한다.

02

데이터 수집

더 많은
한국어 Data set을 수집하고,
인식 범주를 더욱 세분화하여
성능 향상에 주력한다.

03

분석

지속적인 연구와 분석으로
유의미한 결과물을
산출할 수 있도록 한다.

02 새로 만들 SW

01 형태

Single-Page Application

02 구성요소

Input Data : 한국어 문단

Output Data : 해당 문단의 주제(토픽)와 그것의 판단 근거 표시

03 가지고 올 SW

01 TensorFlow

- https://www.tensorflow.org/
- Machine Learning Framework
- 딥러닝 모델 구현에 사용

02 KoNLPy

- https://konlpy-ko.readthedocs.io/ko/v0.5.2/
- 한국어 자연어 처리 Python Library
- 한글 형태소 분석에 사용

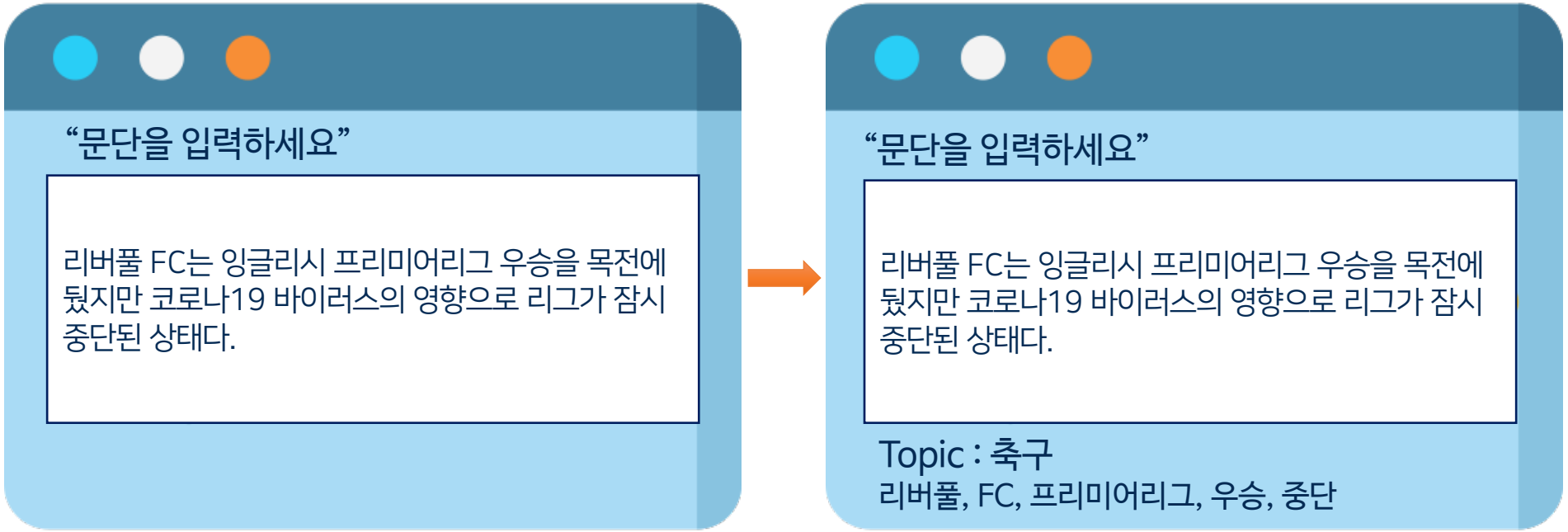
03 React.js

- https://reactjs.org/
- Web Frontend Library
- 결과물을 표시할 Application 제작에 사용

04 Django

- https://www.djangoproject.com/
- Python Web Framework
- 딥러닝 모델을 올려 API화 하는데 사용

04 최종 산출물의 형태 및 기능



05 Alternative Solutions & Project Justifications

01 Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장 (2017.03)

- 유홍연, 고영중

한국어 개체명 인식을 위하여 입력으로 사용되는 단어 표상을 확장하기 위해
사전 학습된 단어 임베딩 벡터, 품사 임베딩 벡터, 음절 기반에서 확장된 단어 임베딩 벡터, 개체명 사전 자질 벡터를 사용했을 때 사전 학습된 단어 임베딩 벡터만 사용한 것 보다 8.05%의 성능 향상을 보인다.

02 효율적 대화 정보 예측을 위한 개체명 인식 연구 (2019. 01)

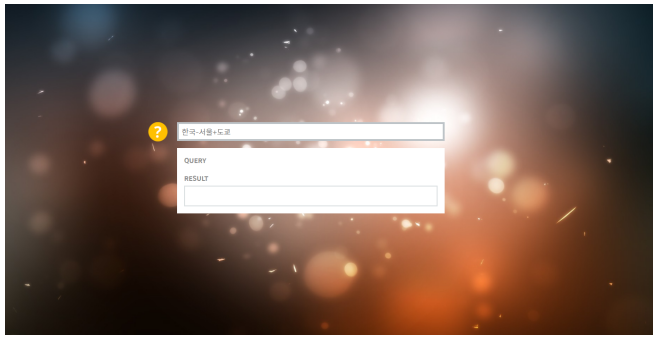
- 고명현, 김학동, 임헌영, 이유림, 지민규, 김원일

효율적인 대화 정보 예측을 위해 전처리 단계에서 사용자 정의 사전을 이용하고, 단어 임베딩 단계에서 최적의 파라미터를 발견한다. 설계한 개체명 인식 모델을 실험하기 위해 생활 화학제품 분야를 선택하고 관련 도메인 내 목적 지향 대화 시스템에서 적용할 수 있는 개체명 인식 모델을 구축한다.

"자체 구축한 대량의 말뭉치를 이용해 단어 임베딩 벡터 학습으로 단어 집합을 확장시켜 기존의 NER보다 성능을 높이도록 한다."

05 Alternative Solutions & Project Justifications

03 word2vector (https://word2vec.kr/search/)



단어 벡터를 이용해 시멘틱 연산이 가능한 사이트로,

고양이+애교=강아지 / 사랑+우정=애교 등의 결과 산출.

"word2vec, 형태소 분석기 이용해,

개체명 인식으로 해당 텍스트 범주를 표시하는 웹사이트 구축"

EX. Input : 손흥민은 대한민국 출신의 축구 선수이다.

Hidden : 손흥민(사람/스포츠), 대한민국(지역/나라),
축구(문명, 문화/스포츠), 선수(문명, 문화/스포츠)

Output : "스포츠"에 관한 문장이다.

06 Risk Analysis & Reduction Plan

01 Risk Analysis

1. 한국어 Data set의 부족
2. 명사 세부 Tag 자체 설정 (ex. 스포츠, 의학, 정치, 경제)
   명사 세부 Tag에 대한 Data set 구축
3. 명사 Tag 및 Data set 자체 설정으로 인한 정확도 예측 불확실성

02 Risk Reduction Plan

1. Korean NER Corpus 등 기존 자료 활용
2. 각 분야(ex. 스포츠, 의학 등) 관련 기사 자료 활용 및 가공해 Data set과 Tag 설정

07 Success Criteria

01 Time

Input Data 입력 후 결과물을 산출하기까지 **“5초 이내”**

02 Success Rate

유의미한 결과물을 산출할 가능성 **“60% 이상”**

감 사 합 니 다