뉴럴 임베딩 기반 주제별 고유명사 연관어 탐색 시스템 개발*

최윤석**, 김한준***
서울시립대학교 전자전기컴퓨터공학부
choiys8819@naver.com**, khj@uos.ac.kr***

Development of neural embedding-based search system for associated proper-noun words

Yoon Seok Choi, Han-joon Kim
School of Electrical and Computer Engineering, University Of Seoul

요 약

연관단어 검색은 질의와 직/간접적으로 관련된 정보들을 쉽게 찾을 수 있게 하고, 새로운 비즈니스 아이디어를 창출하는데 큰 도움을 준다. 기존 검색 포털에서 제공하는 연관단어 검색기능은 주어진 질의어에 대하여 문자열 매칭 수준의 연관단어를 한꺼번에 보여준다. 본 논문은 고유명사 수준의 검색어에 대하여 의미적으로 연관된 다양한 고유명사를 주제별로 분류하여 보여주는 시스템을 소개한다. 본 시스템 구현을 위해서, BERT를 활용한 품사 분석, 개체명 분석 그리고 뉴스 기사를 통해 학습한 단어간 코사인 유사도 분석을 통해, 입력 단어와 의미적으로 연관된 고유명사 단어가 인명, 지명, 기관명 등의 주제로 구분되어 연관도 순으로 출력되도록 한다.

1. 서 론

인터넷 상 방대한 문서 중에서 자신이 원하는 웹문서를 빠르게 확보하기 위한 방법으로서, 질의어와 연관된 단어를 활용하는 방법이 있다. 예를 들어, '손흥민' 이라는 사람을 자세히 파악하기 위해, 이를 검색어로서 질의하였을 때 의미적으로 연관된 다른 축구선수의 이름이나 축구 구단명을 알려준다면, '손흥민' 질의어와 직간접적으로 연계된 유의미한 정보를 신속하게 취득할 수 있을 것이다.

단어간 의미적 연관도를 산정하는 방법으로서, 단어 벡터간 코사인 유사도 분석이 일반적이다. 대개 유사한 의미를 가지는 단어들은 유사한 문맥에서 동시에 출현하는 경향이 크다[1]. 우선 단어들을 벡터공간에 무작위로 매핑하고, 이후 동일한 문장 또는 문서에 동시에 출현하는 단어들이 벡터공간에서 비슷한 방향을 가지도록 학습하여, 의미적으로 연관된 단어벡터들이 가까운 위치에 놓이도록 만들 수 있다. 코사인 유사도 측면에서, 두 단어 벡터의 방향이 동일한 경우 1, 직각을 이루는 경우 0, 완전히 반대 방향인 경우 -1의 값을 가진다. 두 단어에 대한 벡터의 코사인 유사도가 1에 가까울수록, 해당 단어가 동일한 문장 또는 문서에 자주 같이 등장하면서 의미적으로 연관되었음을 말해준다.

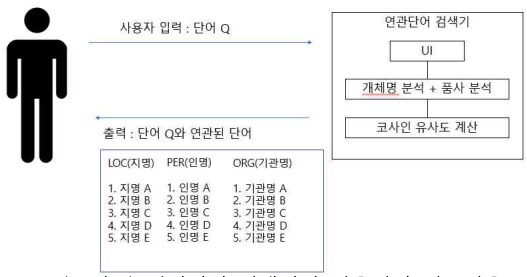
본 논문은 품사(Part-of-Speech) 분석과 개체명(Named Entity) 분석을 통해 고유명사를 인명, 지명, 기관명 등으로 자동 분류하여, 특정 질의어와 연관된 연관 고유명사들을 인명, 지명, 기관명별로 유사도 순에 따라 보여주는 연관단어 시스템을 소개한다.

* 이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원(No. NRF-2022R1A2C1011937, 정형 테이블 데이터셋에 대한 딥러닝 기반 데이터 융합 기술 개발)과 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0-00121, 데이터 품질 평가기반 데이터 고도화 및 데이터셋 보정 기술 개발)을 받아 수행된 연구임.

2. 개체명 연관단어 검색기

2.1 개요

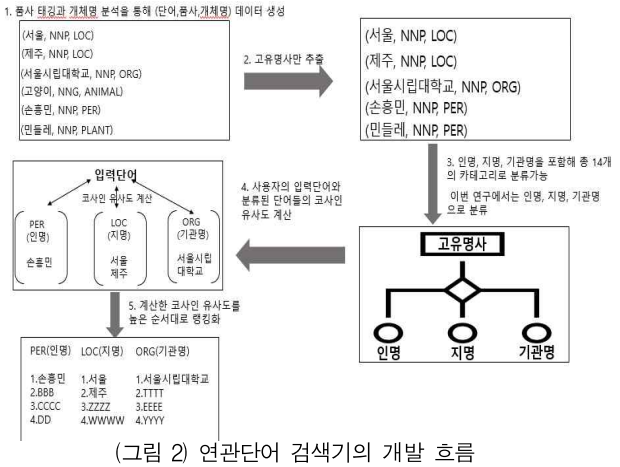
그림 1은 연관단어 검색시스템의 동작 흐름을 보여준다. 사용자가 시스템에 고유명사 단어 Q를 입력하면, 연관단어 검색기는 인명, 지명, 기관명 데이터베이스에 존재하는 모든 단어 벡터와 단어 Q와의 코사인 유사도를 계산하여 상위 N개의 단어를 사용자에게 전달한다.



(그림 1) 연관단어 검색기와 사용자간 상호작용

그림 2는 연관단어 검색기의 내부 구조를 보여준다. 우선 뉴스 기사에 출현하는 단어들에 대한 품사 태깅과 개체명 분석을 통해 각 단어에 대한 품사 및 개체명 정보를 구성한다. 그 중 품사가 고유명사인 단어만 추려내어, 이를 개체명 정보에 따라 인명, 지명, 기관명 등으로 분류한다. 개체명은 인명, 지명, 기관명, 날짜, 숫자, 시간, 동물, 식물, 특정 사건, 의학용어 등, 총 14개의 카테고

리를 포괄하는데, 본 논문에서는 출현빈도가 높은 인명, 지명, 기관명만을 서비스 대상으로 사용한 시스템을 소개한다. 사용자가 질의어를 입력하면, 코사인 유사도 기반 연관도가 높은 고유명사들이 인명, 지명, 기관명별로 출력된다.



(그림 2) 연관단어 검색기의 개발 흐름

2.2 개체명 인식기 및 품사 인식기

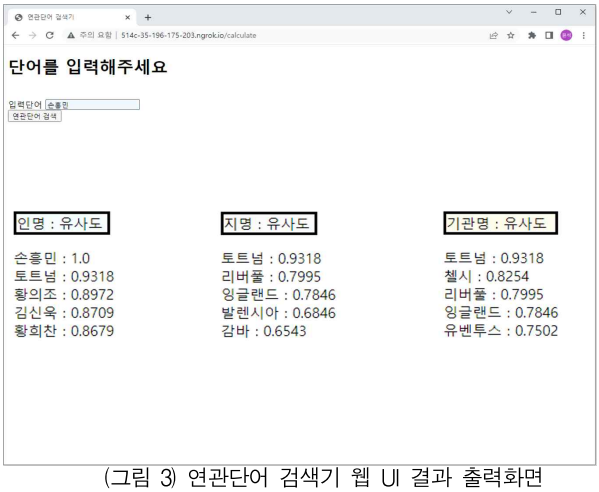
본 시스템의 개발을 위해서, 각 출현단어의 품사와 개체명을 파악하고, 인명, 지명, 기관명 등으로 분류된 고유명사들을 추출할 수 있어야 한다. 개체명 인식 모듈은 네이버 NLP-Challenge 데이터[2]를 BERT에 적용하여 구현된다[3]. 이를 위한 모델 학습은 네이버 NLP-Challenge 내 개체명 데이터를 BERT Tokenizer를 통해 토큰화한 단어와 그에 해당하는 정답 레이블을 BERT에 입력함으로써 수행된다. 개체명 인식 모델의 성능평가 결과, 11,508개 인명, 5,912개 지명, 13,511개 기관명 대한 F1값이 각각 0.83, 0.78, 0.82로 평가되었다. 품사 인식 모델을 구성하기 위해서는 국립국어원의 형태소 분석 말뭉치[4]가 사용되었고, 개체명 인식 모델에서 사용한 모델을 수정하여 구축되었다. 품사 인식 모델의 성능 평가 결과, 고유명사 53,605개에 대한 F1값이 0.96로 평가되었다.

학습이 완료된 모델을 탑재한 개체명 및 품사 모듈은 그것의 출력으로서 토큰화된 단어와 태그를 포함한다. 그런데, 토큰화된 단어가 지나치게 분절되는 경우가 발생하는데, 이 경우 유의미한 어구로 복원하는 작업이 요구된다. 예를 들어, '서울', '##시', '##립', '##대학교'와 같이 세분화된 토큰에 대해서 '서울시립대학교'라는 유의미한 어구로 복원되어야 한다. 이를 위해서 우리는 토큰화된 단어와 함께 출력되는 태그를 이용한다. 즉, ##으로 시작하지 않는 토큰이 들어오면 토큰과 태그를 리스트에 저장해두고, 이후 ##으로 시작하는 토큰이 이어지면 앞서 들어온 토큰에 ##을 제외한 부분을 병합하는 과정을 반복한다. 이후에 ##이 아닌 토큰을 만나면 새로운 어구의 복원작업이 시작된다.

앞서 구축된 개체명, 품사 인식 모델에 BigKinds[5]에서 선정된 1000개의 뉴스 데이터를 적용하여, 출현 단어의 품사와 개체명을 추론하고, 그 중 고유명사이면서 인명, 지명, 기관명인 단어들만 추출한다. 추출한 단어들은 인명, 지명, 기관명 데이터베이스에 분리 저장되고, 향후 사용자 질의단어 벡터와의 코사인 유사도 계산을 수행하여 상위 5개의 연관 고유명사를 출력하게 된다.

2.3 고유명사 연관단어 사용자 인터페이스

그림 3은 고유명사 연관단어 검색기를 위한 웹 기반 사용자 인터페이스를 보여준다. 우리는 이를 구현하기 위해 Python 기반 마이크로 웹 프레임워크인 Flask를 이용하였다. 사용자가 원하는 고유명사 단어(예: 손흥민)를 입력하면, 해당 단어와 의미적으로 연관된 인명, 지명, 기관명별 상위 5개의 연관 단어가 코사인 유사도와 함께 출력된다.



(그림 3) 연관단어 검색기 웹 UI 결과 출력화면

3. 결론

본 논문은 기존 검색포털이 제공하는 연관단어 검색 기능을 개선하기 위해, 주어진 고유명사 질의어와 연관된 고유명사 단어들을 인명, 지명, 기관명별로 분류하는 뉴럴 임베딩 기반 연관단어 탐색 시스템을 소개하였다. 향후 최근 뉴스기사를 크롤링해 자동으로 새로운 단어들을 개체명 데이터베이스에 추가하는 기능을 개발할 예정이다.

참 고 문 헌

[1] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.

[2]https://github.com/naver/nlp-challenge/

[3]https://github.com/kimwoonggon/publicservant_AI/blob/master/5_(BERT_%EC%8B%A4%EC%8A%B5)%ED%95%9C%EA%B5%AD%EC%96%B4_%EA%B0%9C%EC%B2%B4%EB%AA%85_%EC%9D%B8%EC%8B%9D.ipynb

[4]https://corpus.korean.go.kr/request/corpusRegist.do#down

[5]https://www.bigkinds.or.kr/