‘홀로:HolLaw’ 자연어처리(NLP)와 SBERT를 사용한 판례 분석 서비스

윤승현, 김상윤, 이정민, 오지민, 김나연
인천대학교 정보통신공학과
hyeondc99@gmail.com, iamsy703@inu.ac.kr, jeong7162@inu.ac.kr, ojm0623@naver.com, kny010901@inu.ac.kr

‘HolLaw’ A Judicial Precedent Analysis Service using NLP and SBERT

Seung-Hyeon Yoon, Sang-Yoon Kim, Jeong-Min Lee, Ji-Min Oh, Na-Yeon Kim
Dept. of Information and Telecommunication Engineering, Incheon National University

요 약

본 서비스는 문서 내의 가중치를 분석하여 키워드와 관련된 순서대로 정렬하여 판례/법률 검색의 정확도를 향상할 것을 제안한다. 상용화된 다른 판례/법률 관련 서비스의 경우, 키워드 검색을 통해 자신의 사례를 검색할 때, 요약된 정보가 없거나 너무 짧아 사용자가 원하는 판례/법률 결과를 얻을 수가 없어 본 서비스를 기획하게 되었다.

1. 서론

본 논문에서는 TF-IDF 기술이 적용된 자연어처리 알고리즘을 활용하여 판례의 검색과 분석을 통해 쉽게 법률 정보를 얻을 수 있는 판례 분석 서비스를 제안한다. 이를 통해 부당하게 침해받는 자유와 권리를 보장하는 사회 안전망을 구축을 목표로 한다.

본 논문의 구성은 다음과 같다. 본론에서는 구현을 위해 필요한 기술들을 나열하고 구현에서는 서비스의 구성도 및 개발 과정을 설명한다. 결론에서는 추후 서비스가 확장될 경우, 추가로 제공할 수 있는 법률 분야를 서술한다.

2. 본론

2.1 자연어처리(TF-IDF)

자연어처리에서 텍스트를 표현하는 방법으로는 여러 가지 방법이 있다. 그 중 정보 검색과 텍스트 마이닝 분야에서 주로 사용되는 카운트 기반의 텍스트 표현 방법이다.

그중 TF-IDF는 텍스트를 수치화를 하고 나면, 통계적인 접근 방법을 통해 여러 문서로 이루어진 텍스트 데이터가 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내거나, 문서의 핵심어 추출, 검색 엔진에서 검색 결과의 순위 결정, 문서 간의 유사도를 구하는 등의 용도로 사용된다[1]. 문서 내에서 자주 등장하는 단어는 중요도가 낮다고 판단하고, 특정 문서 내에서만 자주 등장하는 단어는 중요도가 높다고 판단한다. TF-IDF 값이 낮으면 중요도가 낮은 것이며, TF-IDF 값이 크면 중요도가 큰 것임을 알 수 있다. 불용어의 경우에는 모든 문서에 자주 등장하기 때문에 불용어의 TF-IDF의 값은 다른 단어의 TF-IDF에 비해서 낮은 값을 나타낸다.

2.2 SBERT - Sentence

Sentence-BERT는 문장 임베딩 성능을 극대화한 모델로써, BERT의 출력 결과에 풀링 연산을 추가한 모델이다. 임베딩 벡터를 표현하고 벡터의 평균값을 풀링하여 벡터를 생성한 후 문장으로 표현할 수 있다. SBERT는 평균 풀링을 사용하고, 평균 풀링을 통해 찾고자 하는 문장을 표현을 얻어 단어의 본질적인 의미를 찾을 수 있다.[2]

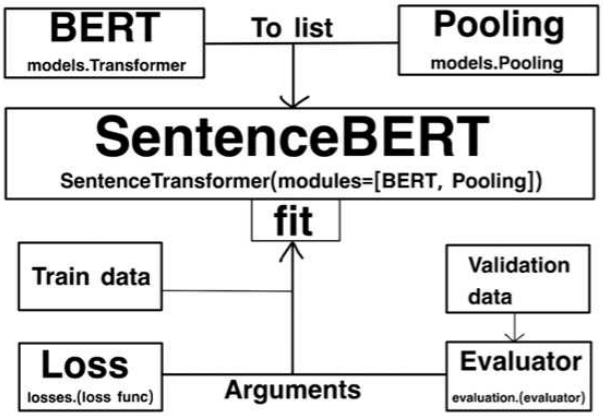


그림 1) SBERT - Sentence

3. 구현

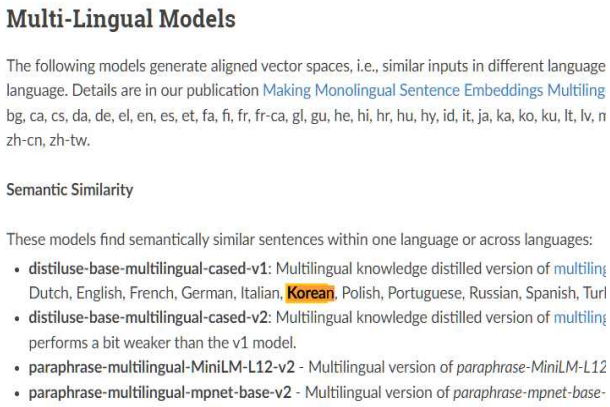
3.1. 데이터

본 연구에서는 국가법령정보센터에 등록된 Open API의 판례 데이터를 수집해 각 내용을 Text 형식으로 구축된 자료를 활용하였다. 약 86,000개의 데이터 중 실제 활용을 위해서 부동산/근로/교통 분야로 한정하여 총 20,000개의 데이터로 축약하였다. 이러한 방식으로 분류된 데이터셋의 80%를 훈련셋, 20%를 테스트셋으로 데이터셋을 구분하였다.

3.2. 데이터 전처리

총 판례 데이터에서 특정 판례에는 1개의 특정 분야의 판례라고 보기 어려운 판례(2개 이상의 분야)로 인해 분류 기준이 모호하여, 판례 데이터를 세 가지 분야로 나누기에 한계가 존재한다. 이를 해결하기 위해 법제처에서 분야별로 분류해 놓은 판례 데이터를 1차로 데이터를 정제하고, 2차로는 분야별 키워드(주택, 임대차)와 같은 관련된 데이터를 추출하여 판례 데이터를 적절하게 선정하여 위 같은 한계를 해결했다.

판례 데이터에는 판시사항, 판결 요지, 참조 조문, 참조판례, 전문 이렇게 구성되고 각 항목에는 법령명 등 숫자와 한자로 구성된 단어가 많고 판사 이름 등 고유명사가 많아 유사도를 검사하기엔 부정확한 신뢰도 문제가 발생한다. 이를 해결하기 위해 각 불용어를 제거하고 TF-IDF와 Knolpy를 활용하여 형태소 토큰화를 수행하고 정리된 단어들을 삽입하여 하나의 문장으로 만들었다. 이를 바탕으로 SentenceTransfomers를 활용하여 신뢰도를 분석한다. SentenceTransformer는 한국어를 지원하기 때문에 문장 유사도를 위해 사용되기 적합하다고 판단하여 신뢰도 검사를 위해 사용했다.



Multi-Lingual Models

The following models generate aligned vector spaces, i.e., similar inputs in different language language. Details are in our publication Making Monolingual Sentence Embeddings Multiling bg, ca, cs, da, de, el, en, es, et, fa, fi, fr, fr-ca, gl, gu, he, hi, hr, hu, hy, id, it, ja, ka, ko, ku, lt, lv, n zh-cn, zh-tw.

Semantic Similarity

These models find semantically similar sentences within one language or across languages:

- **distiluse-base-multilingual-cased-v1**: Multilingual knowledge distilled version of multiling Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, Turl
- **distiluse-base-multilingual-cased-v2**: Multilingual knowledge distilled version of multiling performs a bit weaker than the v1 model.
- **paraphrase-multilingual-MiniLM-L12-v2** - Multilingual version of *paraphrase-MiniLM-L12*
- **paraphrase-multilingual-mpnet-base-v2** - Multilingual version of *paraphrase-mpnet-base-*

그림 2) Sentence Transformers

3.3 데이터 학습 및 결과 분석

사전에 데이터 분석을 끝마친 판례 데이터를 워드 임베딩 벡터값으로 변환하여 사용자의 키워드를 기다린다. 입력 후 키워드를 형태소 단위로 분리한 후 특정 단어를 포함하는 불용어를 제거하고 사전 학습된 SentenceTransformer 모델을 통해 모든 판례 단어들의 워드 임베딩 값에 해당하는 벡터값을 이용하여 하나의 임베딩 벡터로 표현하고 유사도 계산을 통해 입력 키워드와 기존 임베딩 벡터를 비교하여 모든 문서의 신뢰도를 구해, 이후 신뢰도가 높은 순서대로 판례 전문을 사용자에게 제공한다.

```
tensor([[1.0000, 0.6466, 0.7824,  ..., 0.7429, 0.8922, 0.7703],
        [0.6466, 1.0000, 0.4826,  ..., 0.4924, 0.5972, 0.4930],
        [0.7824, 0.4826, 1.0000,  ..., 0.6356, 0.7291, 0.9725],
        ...,
        [0.7429, 0.4924, 0.6356,  ..., 1.0000, 0.9035, 0.6541],
        [0.8922, 0.5972, 0.7291,  ..., 0.9035, 1.0000, 0.7404],
        [0.7703, 0.4930, 0.9725,  ..., 0.6541, 0.7404, 1.0000]])
```

그림 3) 실행 결과 1

실행 결과를 살펴보면 입력값과 가장 유사한 판례 문서에 대한 유사도를 확인할 수 있다. 이를 바탕으로 신뢰도의 값이 큰 판례를 순차적으로 상위에 노출해 사용자가 확인할 수 있다.

3.4 안드로이드 구현

본 서비스는 안드로이드 스튜디오(Android Studio)를 활용하여 안드로이드 애플리케이션을 구현한다.

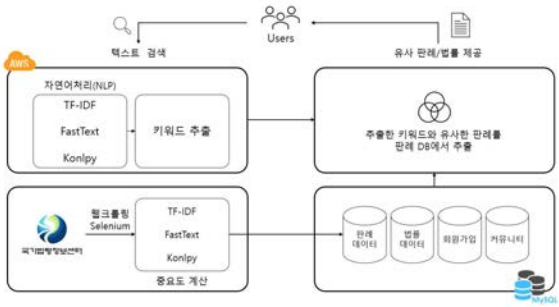


그림 4) 서비스 구성도

3.5 인공지능 챗봇 서비스

본 서비스에서는 대화식 사용자 인터페이스를 모바일 앱에 쉽게 설계하고 통합할 수 있는 자연어 이해 플랫폼으로 Google사의 Dialogflow를 사용한다.

최종 사용자의 표현을 Agent에서 가장 유사한 Intent와 일치시키는 원리로 알맞은 응답을 제공함으로써 실시간 응답 처리가 가능하다[3].

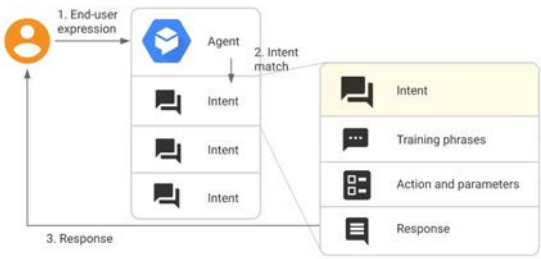


그림 5) Dialogflow 사용자 응답 기본흐름

본 서비스에서는 Dialogflow를 통해 변호사 매칭과 서비스 사용법을 제공함으로써 사용자의 편의성을 높이고자 한다.

3.6 Firebase

커뮤니티 기능과 로그인/회원가입 기능은 Google사의 Firebase를 사용한다.

Firebase Authentication은 Firebase 인증 sdk에 사용자 인증 정보를 전달함으로써 안드로이드 앱 내에서 로그인/회원가입 기능을 제공한다. 인증된 사용자는 서비스의 기능 중 커뮤니티와 변호사 매칭 기능을 사용할 수 있다[4].

Firebase Realtime Database는 클라우드에 호스팅되는 데이터베이스로 데이터는 JSON 트리구조로 저장되고 연결된 모든 클라이언트에 실시간 데이터 동기화 기능을 제공한다[5].

4. 결론 및 향후 연구 방향

4.1 결론

본 서비스는 부동산/임대차, 근로/노동, 교통/운전 분야를 선정하여, 이 주제에 관한 판례와 법률에 대해 높은 접근성과 편의성을 제공하는 것을 목적으로 한다. 단순한 판례/법률 검색이 아닌 관련도순으로 검색하게 하여 사용자가 검색 결과에 만족할 수 있는 서비스를 제공하고자 한다.

4.2 향후 연구 방향

데이터와 분야만 변경하면, 동일한 구조를 바탕으로 확장된 서비스를 제공할 수 있다. 비정형데이터로 이루어진 뉴스나 기사 같은 경우, 특정 단어의 일치만으로 원하는 정보를 파악하기 어려운 점이 있다. 이에 '홀로 알고리즘'으로 문장의 가중치와 유사도를 파악해, 기사나 뉴스 속의 이슈를 파악하는 데 더 나은 정보를 제공할 수 있다. '홀로:판례/법률 분석' 뿐만이 아닌, '홀로:뉴스/기사 분석', '홀로:논문 분석' 등으로 서비스의 확장을 기대할 수 있다.

서비스의 타겟층을 확장하여, 사용자의 스펙트럼을 넓히고자 한다. 또한 부동산/임대차, 근로/노동, 교통/운전 3가지 분야에서 가정법률, 아동 청소년 교육, 정보통신 기술 등의 분야를 추가하여 더 많은 법률 정보를 제공하고자 한다.

※ 본 프로젝트는 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] 유원준/안상준, 딥 러닝을 이용한 자연어 처리 입문, 2022. Won Joon Yoo, Introduction to Deep Learning for Natural Language Processing, Wikidocs

[2] Naoki Shibayama Hiroyuki Shinnou "Construction and Evaluation of Japanese Sentence-BERT Models", 2021

[3] Google Dialogflow
https://cloud.google.com/dialogflow/docs/

[4] Firebase 인증
https://firebase.google.com/docs/auth

[5] Firebase 실시간 데이터베이스
https://firebase.google.com/docs/database