BERT Q&A 모델을 활용한 장학금 정보 추출 및 추천 시스템

강병준[1], 김규진[2], 박진아[3], 장이준[4], 주재현[5], 구형준[6]
성균관대학교 (경영학과[1], 아동청소년학과[2], 컬처앤테크놀로지융합전공[3], 통계학과[4], 경영학과[5]) 학부생
성균관대학교 소프트웨어학 교수[6]

freudbj2@naver.com[1], kyujin000115@gmail.com[2], pja9362@gmail.com[3], jangij0824@gmail.com[4],
dhdktltm117@naver.com[5], kevin.koo@skku.edu[6]

A Recommendation System by Extracting Scholarship Information with a BERT's Q&A Model [1]

Byeongjun Kang[1], Kyujin Kim[2], Jinah Park[3], Ijun Jang[4], Jaehyun Joo[5], Hyungjoon Koo[6]
[1]Dept. of (Business Administration[1], Children and Youth[2], Culture and Technology[3],
Statics[4], Business Administration[5], Computer Science and Engineering[6]) Sungkyunkwan University

요 약

본 논문은 글로벌 이슈로 인한 인플레이션과 대학 등록금 인상 우려 등으로 인해 장학금의 중요성이 부각되고 있는 상황을 고려하여 기존의 장학금 공고 게시물을 수집한 후 BERT Q&A (Bidirectional Encoder Representations from Transformers Question & Answering) 모델을 이용해 개별 맞춤형 장학 공고를 추천하는 시스템을 제안한다. 우선 웹 크롤링을 통해 장학금 정보를 수집하고, BERT Q&A 모델과 사전에 정의한 규칙 기반으로 핵심 정보를 추출한다. 이후 분류 과정을 거쳐 사용자가 입력한 정보와 매칭하여 조건에 맞는 장학금 게시물을 추천할 수 있는 어플리케이션을 구현하였다.

1. 서론

코로나 19 팬데믹과 우크라이나 전쟁 등의 글로벌 이슈로 인해 전 세계적으로 인플레이션이 심화됨에 따라 대학 등록금 동결이 무너질 전망이다[1]. 한국사립대학총장협의회에 따르면 전국 191 개 대학 중 6.3%인 12 개교가 올해 들어 등록금을 인상했으며, 지난 2 월 5 일 한국대학교육협의회 정기총회에 참석한 대학 총장 114 명 중 56 명(49.12%)은 올해나 내년 중 등록금을 인상할 계획이 있다고 밝혔다[2]. 이에 따라 대학생들의 등록금 부담을 완화할 수 있는 장학금의 중요성이 더욱 부각되고 있다.

하지만 성균관대학교를 예로 들면, 학생들은 장학 공고를 잘 확인하지 않는 경향이 있다. 2023 년 3 월 1 일부터 31 일까지 게시된 장학 공고는 총 26 건이며, 해당 공고들의 평균 조회 수는 약 2920 회에 그쳤다. 이는 재학 중인 학부생 25,049 명과 비교했을 때, 중복 조회를 차치해도 평균 11.65%의 학생만 장학 공고를 확인하는 것으로 나타났다.

본 연구에서는 장학금 신청 프로세스에서의 학생들의 부담을 줄여 더 많은 학생들에게 장학금 수혜의 기회를 제공하는 것을 목표로, KorBert(Korean Bert pretrained case)[3] 및 QnA 모델[4]을 활용하여 사용자별 맞춤형 장학 공고를 추천하는 방법을 제안하고 이를 앱으로 구현하였다.

2. 장학금 정보 추출 및 추천 시스템

장학금 게시물 개인화 추천을 위해서는 장학금 게시물 내의 핵심정보 추출하여 사용자의 정보와 비교할 필요가 있다. 이를 위해 성균관대학교 장학금 게시물을 크롤링하고, BERT Q&A 모델과 사전에 정의한 규칙을 통해 게시물 내 핵심정보를 추출하고 카테고리화를 진행하였다. 그 후, 추출된 핵심정보를 사용자가 입력한 정보와 매칭시켜 개인 맞춤형 게시물을 추천하는 과정으로 추천시스템을 구현했다.

2-1. 데이터 수집 및 전처리

학습데이터로는 장학금 공지사항을 크롤링하여 직접 라벨링한 데이터를 사용하였다. 성균관대학교 전체 장학금 공지사항, 단과대학 장학금 공지사항을 크롤링한 이후 KorQuAD 데이터셋[5] 형식을 참고하여 각 게시물에 대해 8 개 질문 (지원가능학년, 최소 학점, 소득 분위, 관련 학과, 거주지, 마감일 등)에 대한 답변을 라벨링하였고, 약 2048 쌍(256 개 게시물 *8 개질문)의 데이터셋을 형성했다. 표 1

[1] 이 논문은 2023년 정부(과학기술정보통신부) 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-01199, 융합보안대학원(성균관대학교))

은 장학금 게시물 데이터셋 예시를 보여준다.

또한, 정답이 존재하지 않는 질문에 대한 답변 라벨링 진행 시 기존과는 다른 라벨링 기준을 채택하였다. 기존 KorQuAD(Korean Question Answering Dataset) 데이터셋에서는 정답이 존재하지 않는 질문에 대한 답변으로 단락 내의 임의의 단어를 집어넣었다. 하지만 장학금 게시물에서 핵심 정보가 없을 때에는 해당 답변이 존재하지 않음을 일관적으로 표시해야 한다. 기존 KorQuAD 데이터셋의 방법을 차용하면 해당 답변이 정답이 존재하는지에 대한 구분이 불가능하기 때문에 매 단락 뒤에 '.'라는 임의의 기호를 붙이고 질문에 대한 답변이 존재하지 않을 시 덧붙인 '.'을 뽑아내도록 데이터 라벨링을 진행하였다.

<표 1> 장학금 게시물 데이터셋

게시물 1	지원가능학년은 무엇입니까?	7 학기
게시물 1	소득분위는 무엇입니까?	5 분위
…	…	…
게시물 2	지원가능학년은 무엇입니까?	신입생
게시물 2	(정답이 존재하지 않는 경우) 소득분위는 무엇입니까?	.

전처리 과정에서 중복되는 게시물과 학습에 유의미하지 않다고 판단되는 게시물은 삭제하고, subword 기반 tokenizing 대신 TTA (Telecommunications Tehcnology Association) 표준 형태소 태그셋을 이용한 tokenizing 을 진행하였다. 형태소를 기준으로 tokenizing 을 진행 시 한국어의 유의미한 의미 단위를 남길 수 있다는 장점이 있다.

2-2. Bert Q&A 모델

MRC(Machine Reading Comprehension; 기계독해), 즉 Q&A 모델은 단락과 임의의 질문이 input 으로 입력되었을 때 output 으로는 그에 대한 답변이 출력되게 하는 모델을 일컫는다. 한국전자통신연구원에서 제공하는 BERT Q&A 사전학습 모델[6]에 새로 생성한 데이터셋을 fine-tuning 하는 방식으로 학습을 진행하였다. 성능 향상 비교를 위해 WordPiece 기반 구글 BERT 모델[7], fine-tuning 전 형태소 기반 표 2 와 같이 ETRI 모델 성능까지 측정하였다.

성능 측정 지표로는 MRC 과제 성능지표로 자주 사용되는 F1 score, EM score 를 사용하였고, 테스트셋으로는 장학금 단락 - 질문 208 쌍을 사용하였다. 하단의 결과는 epochs 5, learning rate 3e-5, batch size 는 4 로 파라미터를 결정하여 측정한 결과값에 해당된다. 최종 모델로는 fine-tuning 후 형태소 기반 ETRI KorBERT 모델을 선정했다.

BERT 모델은 게시물의 일부를 답변으로 추출해내기 때문에 Rule-based 코드를 작성하여 답변을 통일하는 과정을 추가로 진행하였다. 예를 들어, '5 학기 재학생', '3 학년 1 학기 재학생' 모두 '5'라는 숫자로 통일시키는 과정을 일컫는다. 위와 같은 과정으로 추출된 장학금 게시물 핵심정보를 사용자의 정보와 매칭시켜 장학금을 추천해주는 과정으로 전체적인 추천 과정이 진행된다.

<표 2> 모델 성능 측정 결과

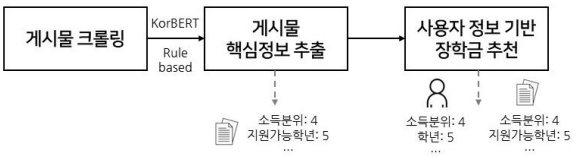
구분	F1	Exact Match
WordPiece 기반 BERT	0.71	0.71
Fine-tuning 전 형태소 기반 KorBERT	0.73	0.74
Fine-tuning 후 형태소 기반 KorBERT	0.76	0.78

3. 어플리케이션 구현 및 결론

본 연구에서는 핵심 정보 추출 및 개인 맞춤형 장학금 추천 서비스를 제공하는 어플리케이션을 개발했다. 이를 위해, 장학 게시글 데이터베이스와 사용자 데이터베이스를 구축하였다. 관계형 데이터베이스를 활용하여, 장학 데이터베이스에는 BERT 학습의 결과인 장학금 핵심정보를 질문에 따라 저장하였다. 사용자 데이터베이스 역시 지원에 필요한 직전 학기 성적, 전체 학기 성적, 소득 분위, 소속 전공, 거주 지역 등의 정보를 저장하였다.

어플리케이션은 <그림 1>과 같이 동작한다. 사용자가 회원가입시 장학금 추천을 위한 기본 정보를 입력한다. 입력된 정보와 데이터베이스에 저장된 장학 정보가 두 개 이상 일치하는 경우, 해당 장학 게시글을 추천 결과로 제공한다. 사용자는 관심 있는 장학 정보를 마이 페이지에 추가 및 삭제할 수 있으며, 추가된 장학의 신청 마감일 알림을 받을 수 있다. 이러한 방식을 통해, 개인화된 장학금 추천을 제공하며 사용자에게 적합한 장학금 지원 기회를 찾는데 도움을 줄 수 있다.

<그림 1> 어플리케이션 동작 방식



참고문헌

[1] 정용택 IBK 투자증권 수석연구위원, "글로벌인플레이션, 펜데믹 대응 과정서 발생한 복합적 현상", KDI 경제정보센터, 2022 년 7 월호

[2] 오종탁 기자, "R 의 공포, 가장 먼저 대학가 덮쳤다", 시사저널, 2023.03.07

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 4171-4186, 2019.

[4] 이동헌. "BERT 를 이용한 한국어 기계독해와 질문생성모델." 국내석사학위논문 강원대학교 대학원, 2021. 강원도

[5] https://korquad.github.io/

[6] https://aiopen.etri.re.kr/bertModel

[7] https://github.com/google-research/bert