SKKULAR : Mobile Application for Scholarship Recommendation

Kyujin Kim[2019310741], Byeongjun Kang[2017314035], Jinah Park[2020314944], Ijun Jang[2019313491], and Jaehyun Joo[2017314245]

Sungkyunkwan University, 2066,
Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do, Republic of Korea
{kyujin0115,freudbj2,pja9362,ijun0824,xhzjtm117}@g.skku.edu

Abstract. This document is a scholarship-matching platform development proposal. The main reason for the design of the platform is to increase students' access to scholarship announcements and to reduce cases where they are unable to apply because they do not know the details of the scholarship. The platform allows clients to receive recommendations for scholarships that meet their conditions by using machine reading AI models, BERT(Bidirectional Encoder Representations from Transformers), and TF-IDF(Term Frequency - Inverse Document Frequency).

Keywords: Scholarship · BERT · TF-IDF.

1 Introduction

Sungkyunkwan University provides information about on-campus and off-campus scholarships periodically in the notice for students. However, our team accidentally discovered that the number of applicants for some scholarships was insufficient. Through the meeting, the causes of the following problems were derived: the above problem occurs because students do not know the scholarship announcements they can apply for. So, we searched for existing services that provide scholarship-related information to students. However, the existing service simply provided information through the scrap function and had limitations in that there was no keyword alarm.

So our team has a goal to reduce the number of cases where students do not even apply for scholarships because they do not know. We designed a platform that has the following functions. First, the application displays the newly updated scholarship, second sends an alarm to appropriate users, third recommends a similar scholarship that the user accessed, and last, sends a notification when the checked scholarship is before the due-date. The service's core technology is: First, we plan to extract keywords from raw data through the BERT model, and second, figure out how the word is important in the document through TF-IDF model.

2 Motivation

One of our team members has worked as an assistant in charge of scholarship for SungKyunKwan University's student aid team for about a year. While working as an assistant, he felt lots of problems with the current scholarship notification and application method of Sungkyunkwan University. One of them was that students were not aware of the scholarship notice. Therefore, for example, there were often situations in which there were five applicants for scholarships in which five students received benefits, or in worse cases, the number of applicants was insufficient compared to the number of beneficiaries. This situation can have the following consequences.

First of all, the burden of work for the department in charge of scholarships increases. If there is a lack of applicants, the staff in charge of the scholarship ask the external scholarship foundation to extend the application deadline. In addition, they would suffer from additional tasks to encourage students who can apply for scholarships through push notifications while searching the school register. If there are no additional applicants, the number of scholarship recipients for the semester will decrease. In addition, next semester's number of scholarship recipients at SungKyunKwan University may decrease when attracting or conducting learnfare or become less persuasive when requesting more beneficiaries, reflecting the lack of applicants last semester.

To prevent the occurrence of these situations and increase students' interest in scholarship notifications, we propose a mobile application called SKKULAR, a customized scholarship notification service, and if possible, we will launch and promote it in collaboration with the student aid team.

3 Problem Statement

The existing process of checking scholarships has the following problems.

3.1 Difficulty of students to check scholarship notices every day

In order to check the newly updated scholarship notices, students must visit SungKyunKwan University's website or each college's website. However, it is difficult for students to visit and check it every day.

Therefore, we will add a function to display the uploaded scholarship notices at once using crawling data so that students can check the scholarship information without visiting the school website every time.

3.2 Difficulty in identifying the qualification requirements

Even if students visit the website every day to check the notices, in order to check the eligibility conditions for application, they should click all notices and check if they are scholarships they can apply for. This unintuitive search process increases students' search fatigue and consequently lowers their interest in scholarship support.

Therefore, we intend to supplement this through the following function. After receiving personal information such as the average grade, income bracket, and the number of semesters in school, the application would send notifications to users who are sufficient to receive scholarships, and the scholarships that students can apply for can be viewed separately.

3.3 Absence of notification functionality

Due to the lack of scholarship-related notification functions, students often have no way of immediately checking newly uploaded scholarship notices, and often forgetting to apply for the scholarship they already checked.

To overcome this inconvenience, we will send a notification when a scholarship notice is posted that students can apply for, and send a notification a week or a day before the deadline to help students prepare in advance and apply for the scholarship.

4 Related Works

There are several services that allow university students to effectively check the notice of scholarships.

4.1 Mobile Application

Uniview is an application that helps students with convenient and smart university life and supports both Android and iOS devices. It shows notices by universities and provides search and alert services by keyword. Key features include: view notices, search and alert by keyword, bookmarks, and quick links to access specific sites directly.

4.2 Website

DreamSpon is a website that provides various scholarship information for university students in real-time. It shows the information on 3,000 scholarships, including scholarships by the school, regional, and foundation scholarships. Furthermore, DreamSpon offers its own scholarships in various forms, including goods, meals, and mentoring, as well as financial support to cheer for university students' dreams. It is also actively using SNS communication channels such as KakaoTalk and Instagram for the efficient operation of the website. Currently, the service is provided in a way that provides information collectively, but the service is also being prepared to provide customized scholarship information.

5 AI-based Background

5.1 Supervised Model - BERT Q&A

In order to gather the necessary information from the scholarship announcement, we will add a dataset labeled in the scholarship announcement to the pre-trained BERT Q&A model for learning. The BERT Q&A model is a supervised model that must understand a given document and question, and find answers within the document.

5.2 Data Set

Dataset would be composed of KorSQUAD's Q&A Dataset [2] and the contents in the scholarship notice. First, the KorSQUAD Q&A Dataset is a Machine Reading Compensation Dataset, a question-and-answer pair for Wikipedia articles. KorSQUAD Dataset has a total of 100,000 pairs of questions and answers.

In addition, the scholarship information dataset is based on the contents of the scholarship notice of Sungkyunkwan University. Here, the announcement of the scholarship of Sungkyunkwan University includes the announcement on the school's official website and the announcement of the scholarship of each department. Scholarship notices are 1,800 on the school's official website and 680 on the college notice. It will also be labeled as household income, residence, year of living, and whether the user gets other scholarships this year, due-date, etc, which are essential requirements for supporting scholarships.

KorSQUAD data is divided into about 80,000 training data and about 20,000 evaluation data. Also, out of the 2,400 scholarship announcement data collected directly, about 2000 documents will be divided into learning data and 400 documents will be divided into evaluation data

5.3 Goal

The goal is to extract core information from the information extracted by rule-based crawling using the pre-trained ETRI model among the BERT Q&A models. For example, if the qualification for a scholarship is "an elementary, middle, high school student, youth, and domestic university student whose parents are living in Suseong-gu for more than one year as of the date of the public announcement", the data should be classified as "parent residence – Suseong-gu, parent residence – one year, job - elementary, middle, high school student, youth, and domestic university student".

Therefore the goal is to extract core information using BERT Q&A, a machine reading model, from the data crawled by the scholarship announcement document.

5.4 Algorithm or Model

The important task of this project is to automatically write data into the Database after extracting key information (ex. eligibility for application (income quantile, grade, credit), deadline and etc) even if any scholarship posts are given as input. This is because such formalized data is required to recommend scholarships that meet the conditions of users. At this time, the NLP model that will be used as a core in the service is the BERT model. In addition, the BERT model will be modified to be applied to MRC(Machine Reading Comprehension) tasks by fine-tuning. It is expected that the BERT model can extract the correct answer to the questions asked by the user and that the desired information can be extracted by asking questions that can extract key information from the scholarship post.

BERT corresponds to a natural language understanding model in which the encoder of the transformer is stacked in a total of 12 layers.

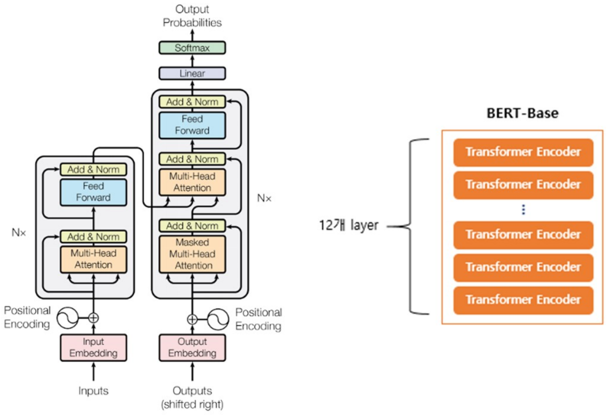

Fig. 1. The operation of BERT model

These BERT has a number of notable features compared to previous models. Transformer consists of an encoder-decoder structure. It can be considered that the encoder creates a context vector and the decoder decodes it. In addition, unlike the previous model Seq2seq, each data is not sequentially input, but data is input at once. The attention method is used to focus on different parts at each point in time to output the results. In other words, the problem of long-term dependency can be solved.

Due to these characteristics, various natural language tasks have become more solvable than before, and tasks that often require logical thinking can be solved in an excellent performance. According to a study on improving Korean MRC performance through transfer learning of pre-trained Korean BERT [4], there are insufficient resources to change the network or the structure of the BERT model itself to do pre-training. Therefore, it is necessary to change the network and show differentiation in other directions rather than restarting pre-training from the beginning.

1) Data Augmentation

In the natural language model, the amount of natural language data is absolutely important. However, natural language data constructed in Korean is far from sufficient, and it is even more difficult to find data specific to a domain. For this reason, it is necessary to label and generate data by hand, but it requires a lot of manpower resources. Therefore, increasing the amount of data by using data augmentation is important. Methods of such data augmentation include Back translation and a method of converting a specific word into a similar word through a Name Entity Recognizer.

2) Hyperparameter tuning

Performance can be improved by tuning the hyperparameter of the BERT model. In the pre-trained model, the number of epochs, Batch Size, Random Seed, and learning Rate, Max Sequence Length can be fine-tuned.

3) Tokenizer [1]

Performance varies depending on Tokenizer. Bert default Tokenizer, especially word piece tokenizer, do BPE (Byte Pair Encoding) by using the likelihood method. It can be said that words are divided and tokenized based on frequency. This technique has the advantage of being able to tokenize words that have not been previously learned as input and are applicable to multiple languages. However, there is a possibility that these techniques may not fully reflect morphemes, an important unit of the Korean language. So ETRI developed a tokenizer that first splits words into morphemes and binds the broken words into tokens through word piece tokenizer to reflect these Korean characteristics well.

4) Knowledge distillation

In the case of a Real-time Service, the model should be driven very fast. At this time, knowledge distillation is needed. It corresponds to the process of learning the loss value that appears to learn the original large model by using it to learn the small model. However, we decided not to adopt this direction because it is not the case that real-time services are required, and the accuracy can be reduced by conducting knowledge distillation.

5.5 Input data and Output data

Data for (body, question) should be input, (answer to question) should be output. The input value enters the model after preprocessing and embedding. In the preprocessing process, the special symbols of the text will be removed and word piece embedding, position embedding, and segment embedding will be performed. The final form is [CLS] paragraph [SEP] correct answer [SEP], and the output will be extracted from the correct part of the paragraph.

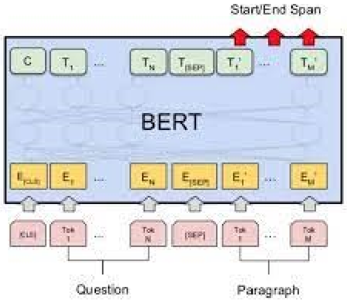

Fig. 2. The format of input data and output data using BERT model

5.6 Existing techniques

In this project, we plan to use a pre-trained model to fine-tuning, create domain-related datasets by hand, or increase performance by changing the tokenizer. At this time, the ETRI pre-trained Bert model and the Korean language specialist tokenizer will be used[3].

5.7 Limitation

The dataset does not include external scholarships. External scholarships are not only wide in scope, but there is a high possibility that the information is distorted. In addition, the attachment would be collected if the file type is PDF. If core information is provided as an image in the scholarship announcement, and if the given PDF file is in image format, the OCR model must be used. We decided not to use an image file and PDF image files because the learning range is very wide. Therefore, the scope of the dataset is included only the college notice and each department notice. In addition, the attachment should not be collected when it is image-type.

5.8 Evaluation

This project will compare the Q&A performance of models trained using fine-tuning methods on pre-trained BERT models. As performance evaluation indicators, F1-score and EM-score will be used to measure the performance of each model. F1-score is the harmonized average value of the precision and reproduction rate between the predicted and actual answers, and EM score is the degree to which the word units of the Korean-based answer exactly match.

6 TF-IDF

6.1 Background

The TF-IDF weight is a method in which a document is expressed as a vector by calculating the importance of words in the document. The words that do not appear frequently in other documents, and that often are written in the document which is calculated can be judged to be of high importance in the document. Therefore, TF means the frequency of each word appearing in a document, IDF is the value taken in the log by taking the reciprocal of how many times a specific word appeared in each document, and TF-IDF is obtained by multiplying TF and IDF values.

Considering TF-IDF, documents can be compared by considering more information than before, and in many cases, they perform better than the Document-Term Matrix (DTM). Therefore, it is mainly used to find the similarity of documents, to determine the importance of search results in search systems, and to find the importance of specific words in documents.

6.2 Goal

In this project, the TF-IDF will be used to find similarities between scholarship documents. TF-IDF, which obtains the importance of words within each document, is expressed as a vector and stored as a value of each document. If the cosine similarity value of the vectors of the two documents is high, it is judged that the two documents are similar. Through this, similar scholarships will be displayed using TF-IDF.

7 Planning in Detail

7.1 Role Distribution

Our project is largely divided into two categories: AI and web app development, and web app can again be divided into front-end and back-end. We decided our main roles based on the interest and competence of each team member. In addition to the main roles of individuals, we will flexibly divide the roles as needed to carry out the project.

Table 1. main roles of each team members

Name	Role
Kyujin Kim	AI, data preprocessing
Ijun Jang	AI, data preprocessing
Jinah Park	Front-end, data collection
Jaehyun Joo	Front-end, data collection
Byeongjun Kang	Back-end, data collection

7.2 Brief Schedule

Based on the roles that were largely divided into two categories, we planned the project schedule by dividing it into AI and app parts. The schedule may vary depending on the situation for more efficient project progress.

Table 2. weekly project schedule

weekly plan							
	2 3 4	5	6 7 8	9 10 11 12	13	14	15
Overall	Project Proposal	Proposal Latex		Midterm		Finalterm	Final Latex
AI		data collection data preprocessing	Rule-Based	Q&A Model (BERT)	Beta Test Debugging		
App		UI/UX design Database setting	frontend work	backend work connect frontend and backend			

References

1. Park, K., Lee, J., Jang, S., & Jung, D. (2020). An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. arXiv. https://doi.org/10.48550/arXiv.2010.02534
2. Korquad. KorQuAD. (n.d.). Retrieved September 30, 2022, from https://korquad.github.io/
3. 공공 인공지능 오픈 api · data 서비스 포털. 공공 인공지능 오픈 API · DATA 서비스 포털. (n.d.). Retrieved September 30, 2022, from `https://aiopen.etri.re.kr/service_dataset.php`
4. 이치훈(Chi Hoon Lee);이연지(Yeon Ji Lee);이동희(Dong Hee Lee). (2020). 사전 학습된 한국어 BERT의 전이학습을 통한 한국어 기계독해 성능개선에 관한 연구. 한국IT서비스학회지, 19(5), 83-91. 10.9716/KITS.2020.19.5.083