Deep Learning Analysis of Korean Certificate Exam Questions

Jinu Kim, Hyeongho Choi, Hakmin Lee
Yonsei University

Abstract

The goal of this project is developing algorithms that can make managing exam data efficiently. Specially, we addressed Korean certification exam data. We developed three models that have different functions each other. In addition to the existing LSTM, we used BERT, a high-performance latest model of natural language processing developed by Google. And we used Konlpy for vectorization of Korea. KoNLpy is a Python library that provides Korean morpheme analysis tools. Lastly, we used different input data to train each model, and we tried to compare functional differences by model and input data. A summary of the model is shown in the figure 1 below.

Keywords: BERT, LSTM, KoNLPy, KoBERT, KLUE

1. Introduction

Figure 1. Model Summary

Model name	Method	Modeling data
Domain of Choice Categorization Model	LSTM	Our data + KoNLPy
Domain of Question Classification Model	Bert	Our data
Similar Question Extraction Model	Bert	External data

Currently, there is a serious educational gap in many areas of education in Korea. Typically, the causes of these problem are regional and economic factors. The reason is as follows.

First, most of the educational infrastructure is concentrated in a specific area, mainly in the metropolitan area. Therefore, students living in non-metropolitan areas have less access to education.

Second, each student has an economic gap. We found that this economic gap is caused by the lack of sufficient quality education services due to cost issues. Statistical indicators of these issues are shown in the figure 2, 3 below.

Figure 2. Comparison of the number of metropolitan, non-metropolitan academies (2022)

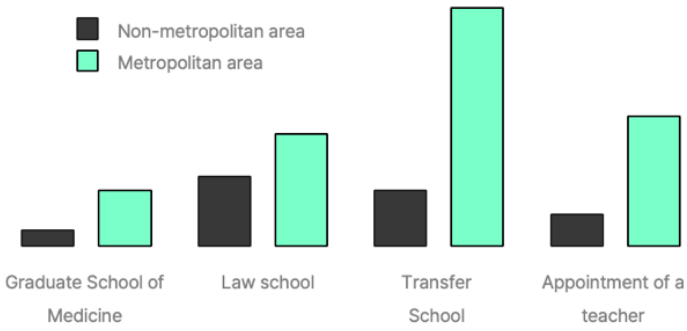
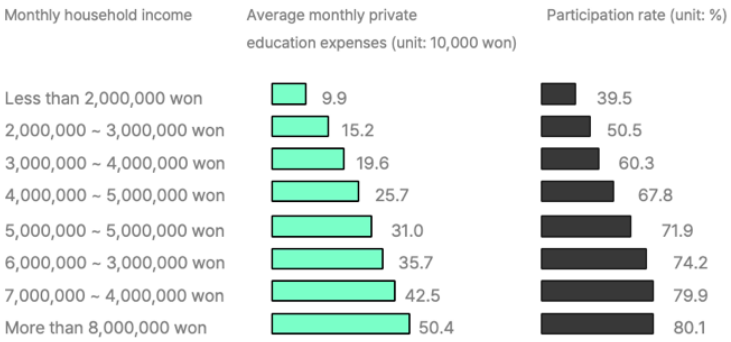


Figure 3. Current status of private education expenses by income class (2020)



In this situation, AI is attracting attention as a technology that can solve these problems. Because AI can overcome regional limitation and provide high-quality education services at a cheap price.

2. Problem Statement

The technical problems that can be solved using these AI technologies are as follows.
So far, we can find the fact that many exam data is not systematically managed in education market. Without organic connection, exist in an analog method. So, without the help of education experts, there is an inefficient in approaching problems that are right for each student.

Therefore, as a solution, we propose an algorithm that connects the relationship between exam data that exist in disordered. We devised four methods to structure exam question data and connect them organically. A description of each method is as follows.

a. Domain of Choice Categorization Model
   - Predicts the exam name based on the choice text
b. Domain of Question Classification Model
   - Predicts exam name based on the question text
c. Question type prediction Model
   - Predicts fields of study based on keywords in the question and choice text.
d. Similar Question Extraction Model
   - Extracts similar problems based on question text.

In conclusion, we succeeded in developing on a, b, d model and failed on c model. The model that predicts question type showed low accuracy. We think it's because of the small amount of data and the keyword extraction method.

3. Data Collection

Our team collected Korea's certification exam data. The data was obtained from the testbank website, the name of q.fran. We used an automatic crawling Google extension software called 'Listly' and collected about 4000 problem data. The raw data is shown in the figure 4 below.

Figure 4. Raw data example

ABEL-1	LABEL-2	LABEL-3	LABEL-4	LABEL-5	LABEL-6	LABEL-7	LABEL-8	LABEL-9
안전관리론								
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	1. 매슬로우(Maslow)의	①	https://q.fran.kr/img/ch	① 안전 욕구	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	2. A사업장의 현황이 다	①	https://q.fran.kr/img/ch	① 0.22	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	3. 보호구 자율안전확인	①	https://q.fran.kr/img/ch	① ㄱ, ㄴ, ㄷ	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	4. 학습지도의 형태 중	①	https://q.fran.kr/img/ch	① 포럼(Forum)	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	5. 보호구 안전인증 고시	①	https://q.fran.kr/img/ch	① #2 ~ #3	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	6. 산업재해의 분석 및	①	https://q.fran.kr/img/ch	① 관리도	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	7. 산업안전보건법령상	①	https://q.fran.kr/img/ch	① ㄱ, ㄴ, ㄷ, ㄹ	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	8. 억측판단이 발생하는	①	https://q.fran.kr/img/ch	① 정보가 불확실할 때	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	9. 하인리히의 사고예방	①	https://q.fran.kr/img/ch	① 사실의 발견	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	9. 하인리히의 사고예방	①	https://q.fran.kr/img/ch	① 사실의 발견	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	11. 산업안전보건법령상	①	https://q.fran.kr/img/ch	① 유해인자의 노출기준	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	12. 버드(Bird)의 재해분	①	https://q.fran.kr/img/ch	① 200	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	13. 산업안전보건법령상	①	https://q.fran.kr/img/ch	① 비계의 조립순서 및	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	14. 산업안전보건법령상	①	https://q.fran.kr/img/ch	① 위험장소 경고	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	15. 학습정도(Level of	①	https://q.fran.kr/img/ch	① 인지 → 이해 → 지각	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	16. 기업 내 정형교육 중	①	https://q.fran.kr/img/ch	① Job Method	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	17. 레빈(Lewin)의 법칙	①	https://q.fran.kr/img/ch	① 행동	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	18. 재해원인을 직접원	①	https://q.fran.kr/img/ch	① 물적 원인	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	19. 산업안전보건법령상	①	https://q.fran.kr/img/ch	① 업무 수행 내용의 기	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	20. 헤드십(headship)의	①	https://q.fran.kr/img/ch	① 지휘형태는 권위주의	②
인간공학 및 시스템안전								
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	21. 위험분석 기법 중 시	①	https://q.fran.kr/img/ch	① PHA	②
	https://q.fran.kr/img/o.	https://q.fran.kr/img/x.p	https://q.fran.kr/img/t.p	22. 상황해석을 잘못하	①	https://q.fran.kr/img/ch	① 실수(Slip)	②

Figure 5. Listly, Google extension software

Figure 6. Q.fran.kr, Testbank website

4. Data Preprocessing

After data collecting, only necessary data was extracted from the raw data. And, we added question keyword column manually. Necessary data was Certificate name, Problem domain, Question text, Choice text and Keywords. Next, missing values are removed. At last, numbers and special characters were eliminated by regular expressions. Through this, everything except Korean, English and spacing were deleted. The preprocessed dataset and column descriptions are shown in the figure 7, table1 each below.

Figure 7. Dataset Example

	A	B	C	D	E	F	G	H	I
1	Exam_name	Question_Type	MainText	Keyword	Choice_Text_1	Choice_Text_2	Choice_Text_3	Choice_Text_4	
2	건축기사	건축계획	1. 기업체가 자사제품의 홍보, 판매 촉진 등을 위해 제품 및 기업에 관한 자료를 소비자들(	전시공간	① 쇼룸	② 런드리	③ 프로시니엄	④ 인포메이션	
3	건축기사	건축계획	2. 사무소 건축의 실단위 계획 중 개실 시스템에 관한 설명으로 옳지 않은 것은?	개실 시스템	① 공사비가 저렴하다.	② 독립성과 쾌적감이 높다.	③ 방길이에 변화를 줄 수 있	④ 방길이에 변화를 줄 수 없다.	
4	건축기사	건축계획	3. 주택단지계획에서 보차분리의 형태 중 평면분리에 해당하지 않는 것은?	동선분리	① T자형	② 루프(loop)	③ 쿨데삭(Cul-de-Sac)	④ 오버브리지(overbridge)	
5	건축기사	건축계획	4. 도서관의 출납 시스템 유형 중 이용자가 자유롭게 도서를 꺼낼 수 있으나 열람석으로	출납 시스템	① 폐가식	② 반개가식	③ 자유개가식	④ 안전개가식	
6	건축기사	건축계획	5. 단독주택에서 다음과 같은 실들을 각각 직상층 및 직하층에 배치할 경우 가장 바람직	단독주택	① 상층: 침실, 하층: 침실	② 상층: 부엌, 하층: 욕실	③ 상층: 욕실, 하층: 침실	④ 상층: 욕실, 하층: 부엌	
7	건축기사	건축계획	6. 다음 중 백화점 매장의 기둥간격 결정 요소와 가장 거리가 먼 것은?	백화점 매장	① 엘리베이터의 배치방법	② 진열장의 치수와 배치방식	③ 지하주차장 주차방식과	④ 층별 매장 구성과 예상 이용 인원	
8	건축기사	건축계획	7. 학교 운영방식에 관한 설명으로 옳지 않은 것은?	학교 운영방식	① 종합교실형은 초등학교	② 교과교실형은 교실의 이	③ 달톤형은 학급과 학년을	④ 플라툰형은 전 학급을 2분단으로 나	
9	건축기사	건축계획	8. 종합병원에서 클로즈드 시스템(closed system)의 외래진료부에 관한 설명으로 옳지	클로즈드 시스템	① 내과는 소규모 진료실을	② 환자의 이용이 편리하도	③ 중앙주사실, 회계, 약국 등	④ 전체병원에 대한 외래진료부의 면적	
10	건축기사	건축계획	9. 공장 건축의 레이아웃(layout)에 관한 설명으로 옳지 않은 것은?	레이아웃	① 제품중심의 데이터보존	② 데이터보존 중앙 유	③ 중앙중심의 데이터보존	④ 고정의 데이터보존 기능이 중첩하	
11	건축기사	건축계획	10. 극장건축의 관련 제실에 관한 설명으로 옳지 않은 것은?	극장건축	① 앤티 룸(anti room)은	② 그린 룸(green room)은	③ 배경제작실의 위치는 무	④ 의상실은 실의 크기가 1인당 최소 8n	
12	건축기사	건축계획	11. 상점의 동선계획에 관한 설명으로 옳지 않은 것은?	동선계획	① 고객동선은 가능한 길게	② 직원동선은 가능한 짧게	③ 상품동선과 직원동선은	④ 고객 출입구와 상품 반입/출 출입구는	
13	건축기사	건축계획	12. 건축공간의 치수계획에서 "압박감을 느끼지 않을 만큼의 천장 높이 결정"은 다음 중	치수계획	① 물리적 스케일	② 생리적 스케일	③ 심리적 스케일	④ 입면적 스케일	
14	건축기사	건축계획	13. 고대 로마 건축물 중 판테온(Pantheon)에 관한 설명으로 옳지 않은 것은?	판테온	① 로툰다 내부는 드럼과 돔	② 직사각형의 입구 공간은	③ 드럼 하부는 깊은 니치와	④ 거대한 돔을 얹은 로툰다와 대형 열주	
15	건축기사	건축계획	14. 극장의 평면형식 중 오픈 스테이지(open stage)형에 관한 설명으로 옳은 것은?	오픈 스테이지	① 연기자가 남측 방향으로	② 강연, 음악회, 독주, 연극	③ 가장 일반적인 극장의 형	④ 무대와 객석이 동일공간에 있는 것으	
16	건축기사	건축계획	15. 다음 설명에 알맞은 사무소 건축의 코어 유형은?	사무소 건축	① 편심형	② 독립형	③ 분리형	④ 중심형	
17	건축기사	건축계획	16. 조선시대에 띠자형 주택으로 대별되는 서민주택의 지방 유형은?	서민 주택	① 서울지방형	② 남부지방형	③ 중부지방형	④ 함경도지방형	
18	건축기사	건축계획	17. 메조넷형(Maisonette Type) 아파트에 관한 설명으로 옳지 않은 것은?	메조넷형	① 설비, 구조적인 해결이 유	② 통로가 없는 층의 평면은	③ 통로가 없는 층의 평면은	④ 엘리베이터 정지층 및 통로 면적의 감	
19	건축기사	건축계획	18. 고딕 성당에 관한 설명으로 옳지 않은 것은?	고딕 성당	① 중앙집중식 배치를 지배	② 건축 형태에서 수직성을	③ 고딕 성당으로는 랭스 성	④ 수평 방향으로 통일되고 연속적인 공	
20	건축기사	건축계획	19. 단독주택의 평면계획에 관한 설명으로 옳지 않은 것은?	단독주택	① 거실은 평면계획상 통로	② 현관의 위치는 대지의 형	③ 부엌은 주택의 서측이나	④ 노인침실은 일조가 충분하고 전망이	
21	건축기사	건축계획	20. 다음 중 호텔의 성격상 연면적에 대한 숙박면적의 비가 가장 큰 것은?	숙박면적	① 리조트 호텔	② 커머셜 호텔	③ 클럽 하우스	④ 레지덴셜 호텔	
22	건축기사	건축시공	21. 벽두께 1.0B, 벽면적 30m2 쌓기에 소요되는 벽돌의 정미량은? (단, 벽돌은 표준형	정미량	① 3900매	② 4095매	③ 4470매	④ 4604매	
23	건축기사	건축시공	22. 석재의 일반적 성질에 관한 설명으로 옳지 않은 것은?	석재	① 석재의 비중은 조암광물	② 석재의 강도에서 인장강	③ 석재의 공극률이 클수록	④ 석재의 강도는 조성결정형이 클수록	
24	건축기사	건축시공	23. Power shovel의 1시간당 추정 굴착 작업량을 다음 조건에 따라 구하면?	굴착 작업량	① 67.2 m3/h	② 74.7 m3/h	③ 82.2 m3/h	④ 89.6 m3/h	
25	건축기사	건축시공	24. 도장작업 시 주의사항으로 옳지 않은 것은?	도장작업	① 도료의 적부를 검토하여	② 도료량을 표준량보다 두	③ 저온 다습 시에는 작업을	④ 피막은 각층마다 충분히 건조 경화한	
26	건축기사	건축시공	25. 콘크리트의 내화, 내열성에 관한 설명으로 옳지 않는 것은?	콘크리트	① 콘크리트의 내화, 내열성	② 콘크리트는 내화성이 우	③ 철근콘크리트 부재의 내	④ 화재를 입은 콘크리트의 탄산화 속도	
27	건축기사	건축시공	26. 아스팔트 방수공사에서 아스팔트 프라이머를 사용하는 가장 중요한 이유는?	아스팔트	① 콘크리트 면의 습기 제거	② 방수층의 습기 침입 방지	③ 콘크리트면과 아스팔트	④ 콘크리트 밀바닥의 균열방지	
28	건축기사	건축시공	27. 콘크리트 배합에 직접적으로 영향을 주는 요소가 아닌 것은?	콘크리트	① 단위수량	② 물-결합재 비	③ 철근의 품질	④ 골재의 입도	
29	건축기사	건축시공	28. 철근, 볼트 등 건축용 강재의 재료시험 항목에서 일반적으로 제외되는 항목은?	건축용 강재	① 압축강도시험	② 인장강도시험	③ 굽힘시험	④ 연신율시험	
30	건축기사	건축시공	29. 발주자에 의한 현장관리로 볼 수 없는 것은?	현장관리	① 착공신고	② 하도급계약	③ 현장회의 운영	④ 클레임 관리	
31	건축기사	건축시공	30. 어스앵커 공법에 관한 설명으로 옳지 않은 것은?	어스앵커	① 버팀대가 없어 굴착공간	② 인접한 구조물의 기초나	③ 대형기계의 반입이 용이	④ 시공 후 검사가 어렵다.	

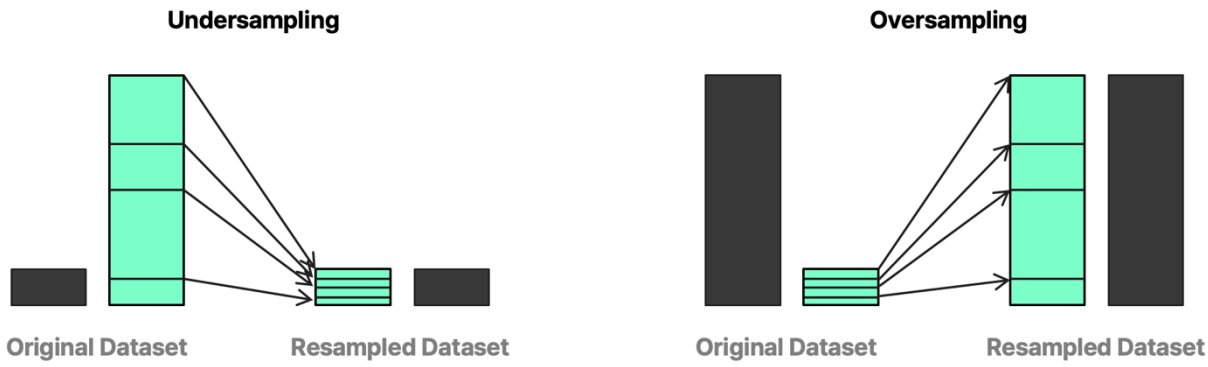
Table 1. Column descriptions

Column Name	Explain
Exam name	Name of certificate exam (5 classes)
Question type	Domain of each question (depend on each test)
Main text	Question sentences
Keyword	Keyword of question sentence (noun)
Choice text	Choice sentences of each question

One of the problems easily encountered when analyzing data is data imbalance. Since securing a lot of data is effective for deep learning analysis, it is recommended to apply an oversampling technique.

One of the factors that degrade the accuracy of the test dataset when modeling artificial intelligence classification is class imbalance. A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. Synthetic Minority Oversampling Technique (SMOTE) was used for data augmentation. SMOTE was used only for train data in the choice text and main text.

Figure 8. Undersampling and Oversampling



Under sampling and over sampling are one of the techniques for solving class imbalances. The SMOTE algorithm is the most widely used model as a method of generating synthetic data among oversampling techniques. SMOTE is a synthetic minority sampling technique that samples multiple classes and interpolates existing minority samples to synthesize new minority instances. This is a method of newly generating data of classes that exist at a low rate using the k-NN algorithm. In conclusion, the result of SMOTE to our dataset is shown in Table2.

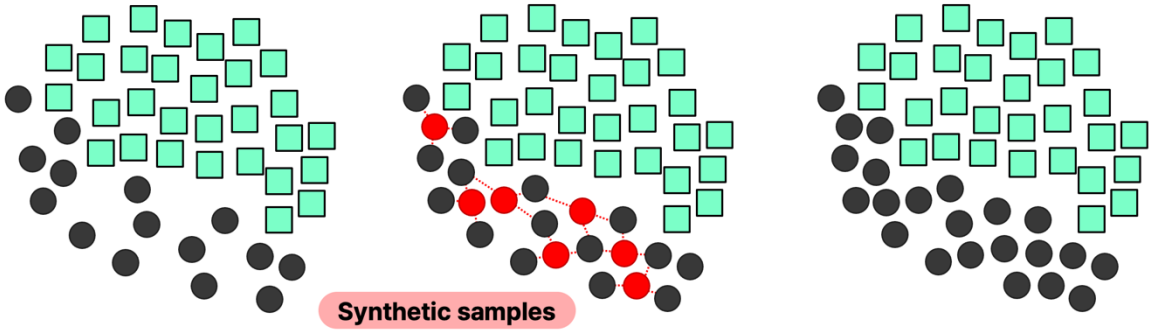
Table 2. SMOTE Results

Label	Number_origin	Number_SMOTE
0	2690	2690
1	2531	2690
2	2155	2690
3	1868	2690
4	1354	2690

In general, it works successfully, but because it works by interpolating between minority data, it reflects only the characteristics between minority data in the modeling set and may be vulnerable to predicting data in new cases. Among oversampling techniques, there is also a method of increasing the amount of data through simple random extraction, and overfitting problems may arise because the data is simply copied. On the other hand, SMOTE generates data based on algorithms, so the probability of overfitting occurrence is less than that of simple random methods. The operation method of SMOTE will be briefly described as follows.

1) Select k neighbor vectors closest to a particular vector among the data in the decimal class
2) Join the reference vector to the selected vector as a line segment
3) Any point on the line segment is a new vector or any one of these.

Figure 9. The process of SMOTE



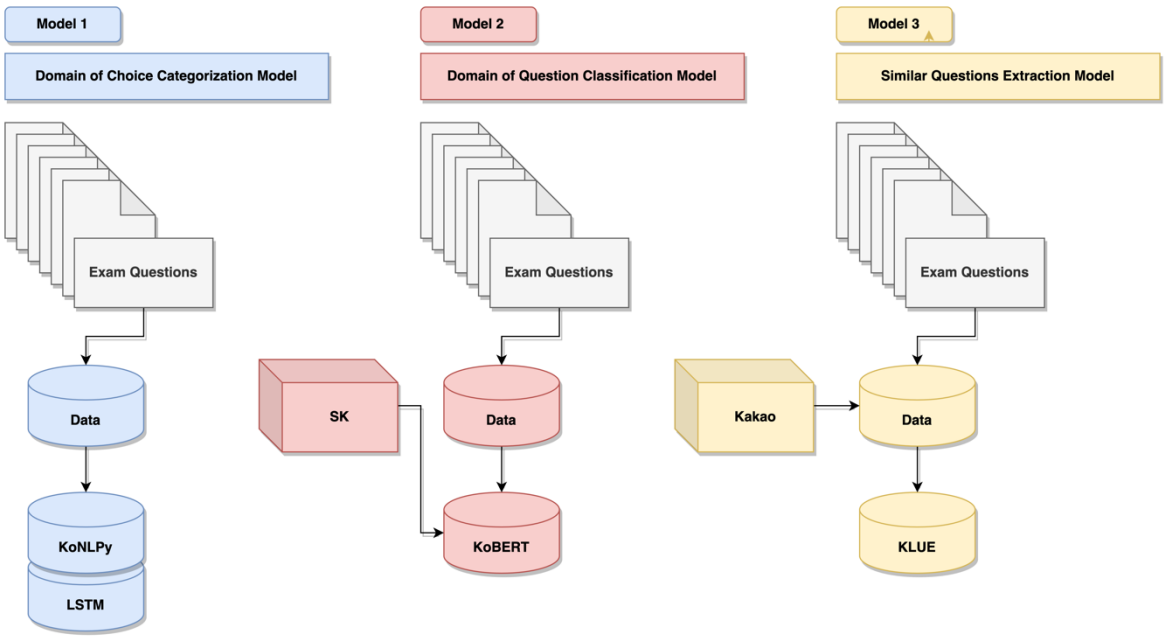
Label	Original Dataset	Generating Samples	Resampled Dataset
0	2690		2690
1	2531	+ 159	2690
2	2155	+ 535	2690
3	1868	+ 822	2690
4	1354	+ 1336	2690

We calculate the difference between a specific vector sample and the nearest neighbor among the decimal data. We have confirmed a maximum of 2690 and a minimum of 1354. Then, multiply this difference by a random number between 0 and 1, and add it to the target vector. The samples that appear to be red are those synthesized using SMOTE.(figure 9) From 0 to 1336 samples were added to each label. Any point can be selected along the line segment between two specific functions. We succeeded in augmenting the data with the largest value as shown in the table above using SMOTE. The final number of resampled datasets is 2690 observed as the maximum value of the original dataset. This is how SMOTE works.

When oversampling is performed, as the amount of data in the class that was a minority is supplemented, a change in precision and reproduction rate occurs. Oversampling should be performed with care so that this change is not too large. As mentioned above, SMOTE is less likely to overfit than a method of randomly performing oversampling.

In addition, there is an advantage that information is not lost and the number of data is not reduced compared to under sampling. However, there are also drawbacks, and from the SMOTE's operating principles, this method samples without considering other classes other than those that exist at low rates. Despite these disadvantages, this project attempted to solve the problem using SMOTE due to its large advantages.

Figure 10. Overall Model Overview



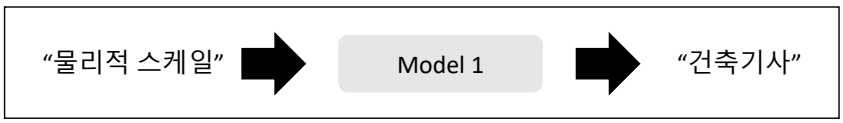
The plot (figure 10) shows the overall model flow for this. In the first model, we tried to infer the exam domain using the choice data for the test questions. We tried to solve the problem by using KoNLPy and LSTM in this model. The second model is a model that attempts to classify domains through questions. It used SK's KoBERT, which showed higher performance than previous models. Finally, the third model was intended to be applied to KLUE using the data provided by Kakao. Through this, we tried to extract similar questions using our question data.

5. Model

5.1 Model 1: Domain of choice categorization model.

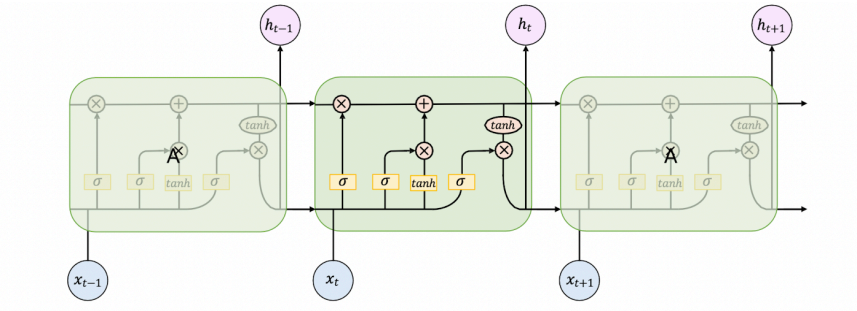
Domain of choice categorization model is a type of NLP classification model. If we input query into this model, we could know a category of the query. Figure 11 is the input and output of the model.

Figure 11. Input and output of the model 1



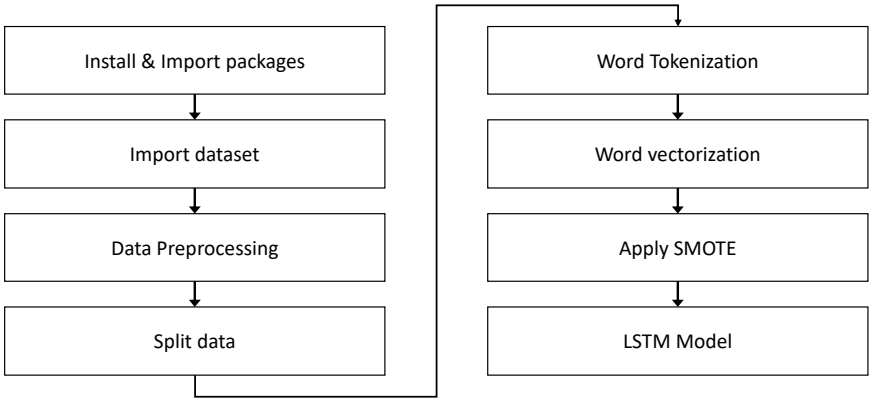
For building model 1, we use LSTM (Long Short-Term Memory) architecture learned in the deep learning class. LSTM is a type of RNN (Recurrent Neural Network) and solves the gradient vanishing problem. Figure 12 is the architecture of LSTM

Figure 12. LSTM architecture



This paper will explain model 1 in order of the flow of code. Figure 13 is the flow chart of model 1 source code. We coded the models in Google Colab. So, first, we install packages that are not in the Colab and import all packages to use. Next, we import the exam dataset excel file for training and testing. With the dataset. We preprocess the data, removing the missing values and everything except Korean, English, and number in the data. Also, we change the y label from string to integer and split the data for the training and testing.

Figure 13. Flow char of model 1 source code



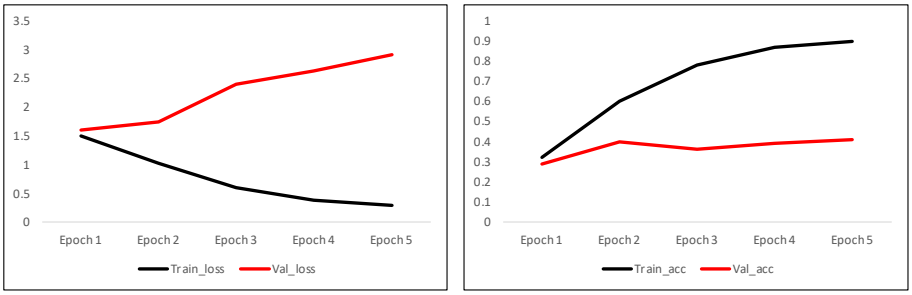
As the beginning of the word embedding process, we use Okt function of KoNLpy packages to tokenize the sentence. With the token, we do word embedding by using tensorflow integer indexing vectorization function. We make the length of all vectors to 30. After the

word embedding, we use SMOTE to solve the data imbalance and transform the y variable to dummy variable.

Next step is building the LSTM layers. We use embedding layer in first layer and 5 node in last layer for classifying 5 class. We use default activation and recurrent activation in LSTM layer and softmax activation in last layer. After the setting, we train the model using training data for validation and using early stopping option.

Figure 14 is the results of model1. Validation loss of model 1 increase quickly, so training step stop early. To solve this problem, we tried to simplify layers, use various dropout layer, data augmentation, etc. But we failed to solve the problem and conclude it as insufficient data problem frequently occurring in NLP

Figure 14. Result of model 1



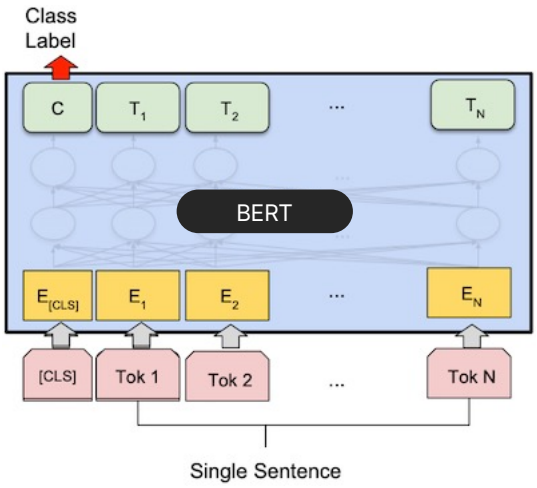

```
Epoch 1/10
470/471 [============================>.] - ETA: 0s - loss: 1.4127 - acc: 0.3994
Epoch 1: val_acc improved from -inf to 0.30632, saving model to best_model.h5
471/471 [==============================] - 13s 10ms/step - loss: 1.4122 - acc: 0.3999 - val_loss: 1.7860 - val_acc: 0.306
Epoch 2/10
464/471 [============================>.] - ETA: 0s - loss: 0.7679 - acc: 0.7172
Epoch 2: val_acc improved from 0.30632 to 0.40768, saving model to best_model.h5
471/471 [==============================] - 3s 7ms/step - loss: 0.7653 - acc: 0.7182 - val_loss: 1.9240 - val_acc: 0.4077
Epoch 3/10
471/471 [==============================] - ETA: 0s - loss: 0.4105 - acc: 0.8588
Epoch 3: val_acc did not improve from 0.40768
471/471 [==============================] - 3s 7ms/step - loss: 0.4105 - acc: 0.8588 - val_loss: 2.2011 - val_acc: 0.4000
Epoch 4/10
468/471 [============================>.] - ETA: 0s - loss: 0.2656 - acc: 0.9056
Epoch 4: val_acc did not improve from 0.40768
471/471 [==============================] - 3s 7ms/step - loss: 0.2668 - acc: 0.9055 - val_loss: 2.6441 - val_acc: 0.4072
Epoch 5/10
470/471 [============================>.] - ETA: 0s - loss: 0.2157 - acc: 0.9198
Epoch 5: val_acc did not improve from 0.40768
471/471 [==============================] - 3s 7ms/step - loss: 0.2157 - acc: 0.9198 - val_loss: 2.8441 - val_acc: 0.3990
Epoch 5: early stopping
```

Test Acc: 0.6944

5.2 Model 2: Domain of Question Classification Model

To solve data problem appearing in model 1, we use the BERT, which is pre-trained NLP model, in model 2 and model 3. BERT is a transformer-based model for NLP developed by Google. Because original BERT learned only English data, we use KoBERT, which is Korean BERT model, developed by SK. It also provides word embedding model. Figure 15 is the architecture of BERT and KoBERT model summary.

Figure 15. BERT architecture and KoBERT model summary



TrainSet

Data source	Sentence	Token
Korean WIKI	5M	54M

Vocabulary

- Size : 8,002
- SentencePiece: Korean WIK based Tokenizer
- Less number of parameters

Like model 1. we will explain model 2 in order of the flow of code. Figure 16 is the flow chart of model 2 source code. Firstly, we install and import required packages and our data, and preprocess the data like model 1. After preprocessing, we split the data for training and testing and do word embedding by using KoBERT tokenizer function.
We load the pre-trained KoBERT model and set the hyperparameters for fine-tuning. After setting, we do fine-tune the KoBERT model for our purpose by using our own dataset.

Figure 16. Flow char of model 2 source code

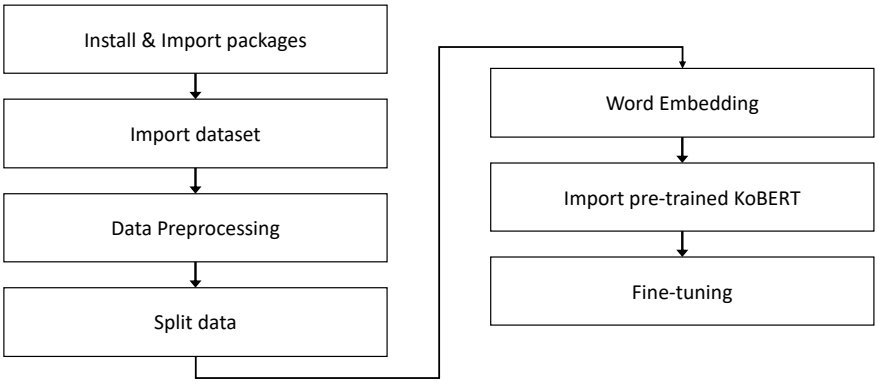
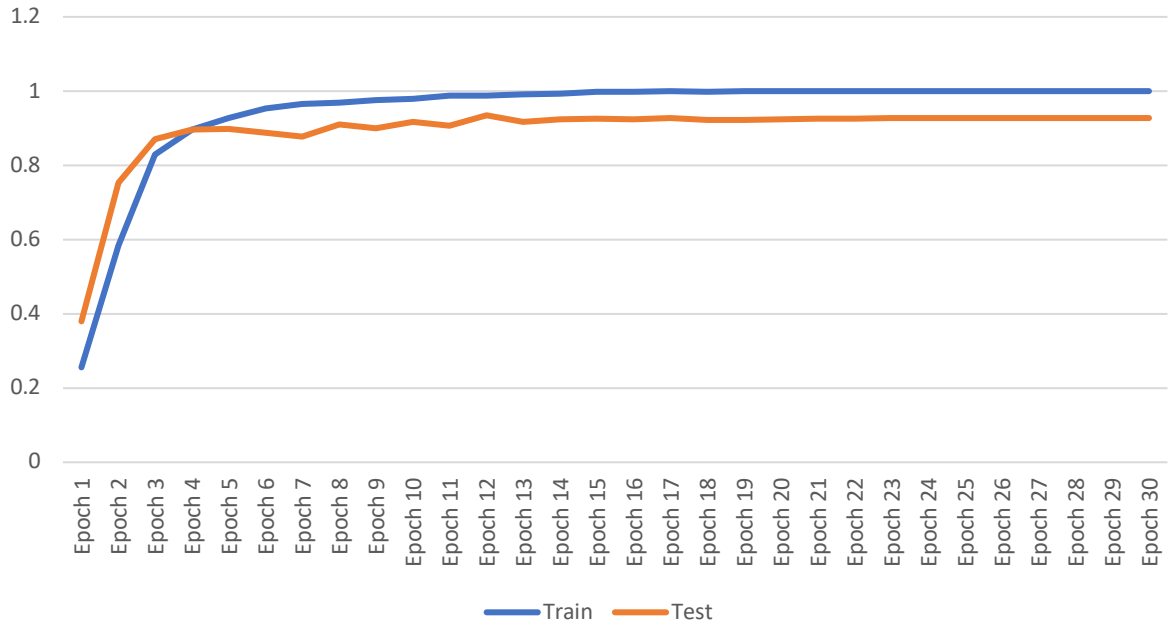


Figure 17 is the results of model 2. Performance of model 2 is well to classify the query. When the epoch approach 12, the accuracy of the train data exceeded 0.93.

Figure 17. Performance of model 2



5.3 Model 3: Similar Questions Extraction Model

To make similar question extraction model, we use KLUE model, which is pre-trained model specialized in sentence similarity, and use Kakao open dataset. Table is the example of Kakao dataset. We constructed the model 3 according to the KLUE guideline.

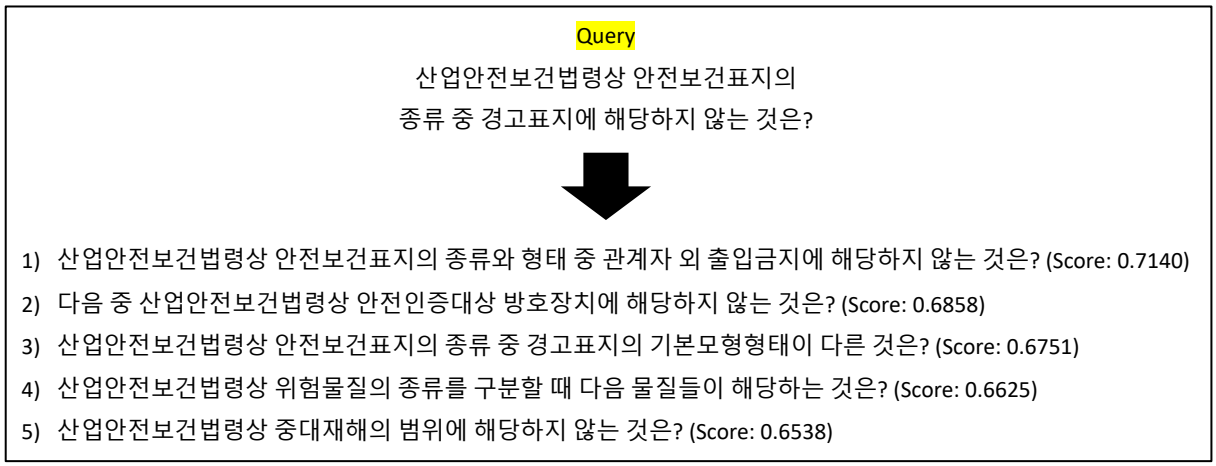
Table 3. Sample of Kakao dataset

id	score	sentence1	sentence2
1	5.000	비행기가 이륙하고 있다.	비행기가 이륙하고 있다.
4	3.800	한 남자가 큰 플루트를 연주하고 있다.	남자가 플루트를 연주하고 있다.
5	3.800	한 남자가 피자에 치즈를 뿌려놓고 있다.	한 남자가 구운 피자에 치즈 조각을 뿌려놓고 있다.

First, we install required package, and import the packages, KLUE pre-trained model, embedding model, and Kakao dataset. Following the guideline, we fine-tune the KLUE model.

To extract similar question by using model 3, we preprocess our own dataset and input these into model 3. Figure 17 is the result of model 3. If we input question query into the model 3, we could be returned similar question sentences and scores in dataset.

Figure 17. result of model 3



6. Conclusion

We built three NLP models for different purposes with Korean Certificate Exam Questions. As a result, we used both our own LSTM model and the latest developed BERT model. Also, we used not only our own crawled dataset but also Kakao open dataset. This study contributes to use NLP models to the new domain and would serve as a backbone for the next study. We expect this experience to help people in education industry and education policymakers.

References

Myeong-Cheol Jwa, Jeong-Woo Jwa. (2022). Development of Tourism Information Named Entity Recognition Datasets for the Fine-tune KoBERT-CRF Model. International Journal of Internet, Broadcasting and Communication, 14(2), 55-62.

Jwa, M.-C., & Jwa, J.-W. (2022). Development of Tourism Information Named Entity Recognition Datasets for the Fine-tune KoBERT-CRF Model. International Journal of Internet, Broadcasting and Communication, 14(2), 55–62. https://doi.org/10.7236/IJIBC.2022.14.2.55

Hyunji Kim. (2021). A Study on Brand Image Analysis of Gaming Business Corporation using KoBERT and Twitter Data. Journal of Korea Game Society, 21(6), 75-85.

Hannah Lee, Hyewon Bae, Kyuhong Shim, Wonyong Sung. (2022). Four Character Idiom Classification using KoBERT. Journal of the Korean Society of Information Science, 2129-2131.

KyuHwon Park, Young-Seob Jeong. (2021). Korean Daily Conversation Topics Classification Using KoBERT. Journal of the Korean Society of Information Science, 1735-1737.

A-Gyeong Kim, Young-Seob Jeong. (2021). Topic classification of domestic music using KoBERT. Journal of the Korean Society of Information Science, 1738-1740.

Park S, Moon J, Kim S, et al. KLUE: Korean Language Understanding Evaluation. arXivorg. Published online 2021. doi:10.48550/arXiv.2105.09680

SKTBrain. KoBERT. GitHub. https://github.com/SKTBrain/KoBERT/tree/master/kobert_hf

KoBERT. (2021, February 14). SKT Open Source. https://sktelecom.github.io/project/kobert/

Text classification using RNN and CNN. Gist. https://gist.github.com/Lucia-KIM/165b8f13c007f83b4762ab436ea95610

Huffon. Huffon/klue-transformers-tutorial. GitHub. https://github.com/Huffon/klue-transformers-tutorial

KLUE-benchmark. Korean NLU Benchmark. GitHub. https://github.com/KLUE-benchmark/KLUE