Research Article

수능 영어시험의 어휘 등장 패턴 분석

반병현[1], 이제종[2], 황현목[3]

[123]프로젝트 균형감각

요 약

텍스트 마이닝 기법 기반의 어휘 분류 기법과 LSTM을 활용한 어휘 출현 패턴 예측 방법을 소개한다. 단순한 텍스트의 등장 빈도를 기반으로 한 프리프로세싱 기법을 제안하며, 이 기법으로 제작된 어휘 스크리닝 기술은 시중의 베스트셀러 수능 어휘 교재인 *워드마스터* 대비 2023년도 수능 시험 적중률이 4.35배 더 높았다. LSTM 딥러닝 기법을 활용한 인공지능 기술도 함께 제안하며, 이 기법은 수능 시험에서 특정 어휘가 어떤 패턴으로 등장하는지 분석하고 이를 토대로 다음 수능에 출제될 어휘의 목록과 그 등장확률을 예측할 수 있다. AI의 예측값은 여러 크기의 데이터 윈도우를 바탕으로 학습된 추론값을 노멀라이즈한 뒤 가중합 연산하여 구한 것이며, 저자들은 이 값을 AI-Score라고 부른다. AI-score는 특정 단어가 다음 수능 시험에서 등장할 확률을 의미하며, 논문에서 제안하는 방법론이 100점으로 분류한 단어들은 2023수능에서 100% 출제되었다. AI-Score가 60점 이상인 영역에서는 불과 1.7%수준의 오차로 수능에 출제될 단어들의 등장 확률을 예측하는 데 성공했다.

▪ 중심어 : 자연어처리, 텍스트 마이닝, 딥 러닝, 패턴 예측, 수능

Analysis on English Vocabulary Appearance Pattern in Korean CSAT

Byunghyun Ban[1], Jejong Lee[2] and Hyeonmok Hwang[3]

[123]Proj. Sense of Balance

ABSTRACT

A text-mining-based word class categorization method and LSTM-based vocabulary pattern prediction method are introduced in this paper. A preprocessing method based on simple text appearance frequency analysis is first described. This method was developed as a data screening tool but showed 4.35 times higher prediction accuracy compared to Word Master book. An LSTM deep learning method is also suggested for vocabulary appearance pattern prediction method. AI performs a regression with various size of data window of previous exams to predict the probabilities of word appearance in the next exam. Predicted values of AI over various data windows are processed into a single score as a weighted sum, which we call an "AI-Score", which represents the probability of word appearance in next year's exam. Suggested method showed 100% accuracy at the range 100-score area and showed only 1.7% error of prediction in the section where the scores were over 60 points.

▪ Keyword : Natural language processing, Text mining, Deep learning, Pattern prediction, CSAT

I. Introduction

Korean CSAT(대학수학능력시험) is an annual test administered by Korean Ministry of Education. Every Korean student takes K-CSAT for admission of university. As K-CSAT is held only once a year, Korea Institute of Curriculum and Evaluation (KICE) provides official mock exams in June and September. On average, 510 thousand students take K-CSAT every year [1]. As previous questions from K-CSAT are freely available at KICE, many instructors and educational institutes struggle for analyzing the pattern itself and the change of patterns over year from previous exams.

This paper provides a prediction matehod for english vocabulary appearance pattern with big-data mining and deep learning.

Previous works related to exam pattern analysis and natural language processing are summarized on II. Related works.
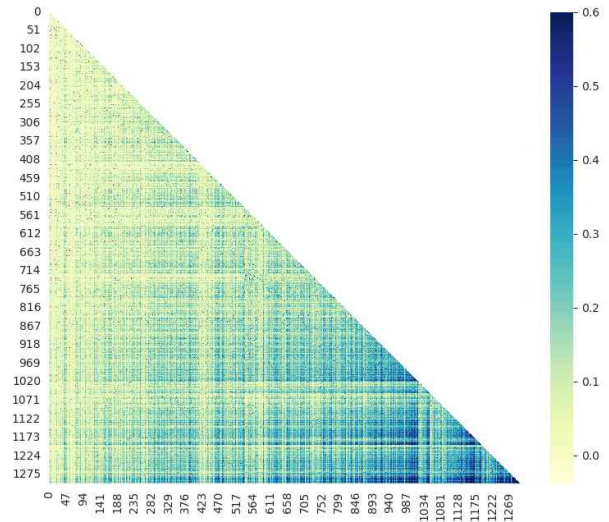
Details on data acquisition, preprocessing, screening and AI-based prediction are described in III. Method. It also provides detailed training strategies such as data windowing and loss function, and normalization method to convert AI-predicted logit into a score which measures the probability of word appearance.

IV. Results. shows the distribution and accuracy of AI-predicted values, and the prediction accuracy on 2023 K-CSAT English exam.

II. Related Works

1. Patterns of Exams

The author of Word Master introduces that he gathered English words from K-CSAT, official mock exam of K-CSAT, government authorized textbooks to write the vocabulary book [2]. Seonsik Kyeong claims that his book, which contains 4,138 words, predicted English vocabulary of K-CSAT with 93.3% ~ 98.5%

[그림 1] 문장 내 단어 상관관계 히트맵
[Fig. 1] Heatmap of Word Correlation in Sentences

accuracy for recent 3 years [3]. But each year's exam contained less than 1,600 words so maximum possible accuracy cannot numerically exceed 38.7%, which is achieved only when every word from an exam is contained in the book. It seems that the author did not apply mathematical definition: number of correct predictions over total number of predictions, or the result is not accurate.

Those two books are most popular in Korean English education book market, but not any statistical analysis results or big data analysis method about them are available.

Hyerang Lim provided a prediction method of difficulty of Korean CSAT English test, based on corpus analysis [4]. She applied Lexical Complexity Analyzer and L2 Syntactic Complexity Analyzer to investigate the relationship between linguistic features and item difficulty of a Mock CSAT English test. Y. Wang, et al. [5] and C. Zhang, et al. [6] proposed a neural network-based model with Long Short-Term Memory(LSTM) for SAT problem solving. Their LSTM model learns the question-answering pattern from SAT

exam to perform problem solving.

Recent studies even provide various generative methods for educational-purposed question generation with artificial neural networks [6-9]. As even various generative model for question-answering is available, pattern analysis of a serial exam is not difficult at all.

2. Natural Language Processing

Natural Language Processing (NLP) is a computer-science area for understanding and generating the patterns of human-speaking languages. Although linguistic definition of natural language is limited as naturally formed language along a community, NLP technologies can be also applied to artificial languages such as Esperanto or programming languages.

NLP approaches generally use linear algebraic approaches. So, most approaches for text NLP first convert the text or spoken words into tensor of integers. One of the easiest but most effective way is word vector embedding. Word vector embedding first requires a table of words to convert each word into a number. Whole vocabularies for analysis should be enlisted on the table. After writing the table, input sentences are tokenized and converted into a series of number, referencing the table. This method is called word2vec [10]. A sparse representation of word vectors, which just convert the word embedding vector into a set of one-hot vectors, can be applied for easier classification of each word because the one-hot vector representation is described as 2D data array which fits best for linear algebraic methods and neural networks.

Long Short-Term Memory(LSTM) was proposed by S.Hochreiter in 1997. [11] As LSTM has dynamic forget gate architecture, it has great advantage in learning time-series data and languages. Although transformer or CNN methods for NLP are highly recognized today,

LSTM based NLP models are still the most popular algorithms [12].

III. Method

1. Data

KICE's website (https://www.kice.re.kr) provides K-CSAT exam PDF files and the official mock test files. We obtained K-CSAT test files from 2003 to 2022, and mock test from 2004 Jun to 2022 September. As there were two types of English exam in 2014, total 21 PDF files of K-CSAT English exam and 40 mock test files were downloaded.

We used K-CSAT exam files for main analysis and the training resources of AI, while we used the mock test files for optimization of our method but not for training the neural networks.
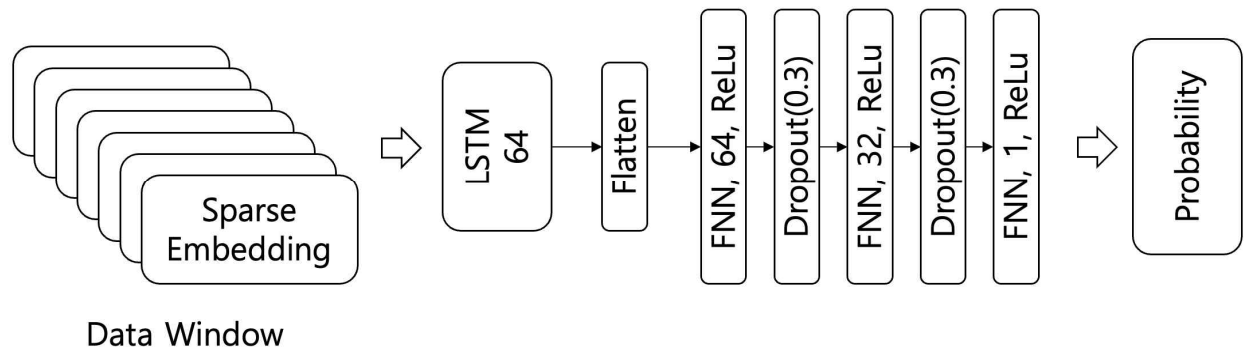
2. Preprocessing

All PDF files were converted into txt files with Python codes (Supplement Data 1). We have removed all Korean words and escape characters. We also replaced any useless patterns, Unicode errors, and some string patterns which appear only on some specific year's exam. 20 samples of removed tokens are described on Table 1.

<표 1> 제거된 토큰 샘플 20건
<Table 1> 20 Samples of Removed Tokens

"\U000f0802"	" \n. \n"	"(A) , "	"fffbg.\n"
"\U000f003b"	". ,"	"(B) , "	"(e.\ng"
"\x0c"	"'''"	"(C) , "	" a.\nm."
"\xa0"	"_____"	"(D) , "	" p.\nm."
"/return"	"--"	"(E) , "	"'''"

The full list of removed tokens are indicated on the read_sunung_txts() function of our source code (Supplement Data 2). After preprocessing, we also

[그림 2] 뉴럴 네트워크 구조
[Fig. 2] Neural Network Architecture

extracted sentences for analysis of word correlation in sentences. Then the processed texts are tokenized for frequency analysis.

3. Frequency Analysis

6,575 words have appeared in K-CSAT, and 6,951 words have appeared in the mock test. 2,262 words have appeared in the mock test but never have appeared in the official K-CSAT. Also 1,886 words have only appeared in the official K-CSAT but never in the mock test. 4,689 words have appeared in both exams. The result shows that vocabulary tendency of the two tests is different.

However, word tokens were not directly used for pattern analysis because raw tokens contain not only the original forms of words but also plural forms of nouns, third person singular form of verbs, past tense form of verbs, and past or present participle form of verbs, etc. Therefore another step for restoring the original forms of word tokens were required. So we have built a dictionary which maps the words into their original forms with Selenium and Chrome Driver to convert the word tokens into their original form. Then the appearance frequency of each word was calculated along years.

4. Word Screening by Experts

It is not so recommended to sort whole words by the frequencies of appearance because the grammatically repeated words, articles, pronouns, "to" for to infinitive, and many other tokens should be highly-ranked on the list. To analyze vocabulary tendency of time-series document data, we first had to remove some words which just technically have high-appearance-frequencies but do not provide patterns underlying sequential K-CSAT exam.

To remove such artifact for higher-resolution of analysis, two English education experts participated in the screening work; one was a public education teacher and the other was a private education teacher. They have screened proper nouns, inappropriate tokens, conjugations of verbs which computer could not process, and other words which do not fully reflect K-CSAT exam tendencies. After screening process, only 2,500 words were survived. The words were then categorized into two groups by frequencies of appearance.

Word group of 'High rank' contained 1,300 words, and the other group had 1,200 words. Finally, we used the 1,300 word group for training of AI.

The correlation of tokens of 'High rank' group in

sentences are described in Fig 1. It shows that only small rate of words shows high correlation. We consider this result shows the removal of artifacts, because the result shows sparse correlation between words, which implies generalization of source data.

The frequency analysis result for 9,329 tokenized words is available at the authors' repository (Supplement Data 4). And 1,300 words' data over 19 years are available (Supplement Data 5).
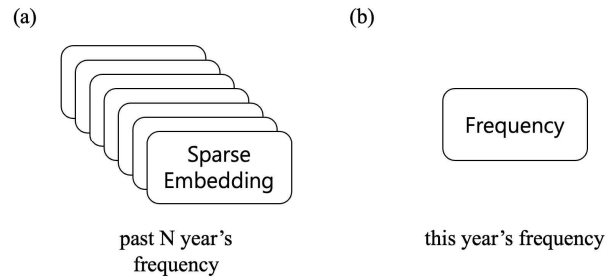
5. Neural Network Architecture

We designed a small deep neural network model because the time-scale dimension of data was too small. As K-CSAT exam has only 20 years of history, a big model may show severe overfitting.

The whole architecture is displayed on Fig 2. The model has single LSTM layer and 3 FNN layers. Flatten layer is applied to convert the shape of LSTM output into lower dimension tensor. First 2 FNN layers has dropout with rate 0.3. The output layer does not have a softMax or other nonlinear activation function because the post-processing step just can normalize the positive values of output logits. ReLu was applied to remove negative values because negative-signed probability is not needed to predict word appearance pattern. The source code is uploaded online (Supplement Data 3).

6. Data Windowing

In order to learn the previous exam and to predict the appearance probability of future exam, we used data windowing strategy. Windowing strategy is described in Fig 3. Each window consists of past N year's frequency tensor data. As a frequency distribution of 1,300 words in each year is a vector, 1D data, the elongation of data on the perpendicular

(a)　　　　　　　　(b)

Sparse Embedding

Frequency

past N year's frequency

this year's frequency

[그림 3] 데이터 윈도잉 전략
[Fig. 3] Data Windowing Strategy

direction forms 2D matrix.

A minibatch for training obtains an array of several windows, therefore, the input LSTM layer of the neural network receives a (1300, 1, N, B) size tensor, where B is the size of minibatch. Or it can be also represented into (1300, N, B) by squeezing the unnecessary dimension.

7. Training

The model is then trained with the loss function described in equation (1). $X_{t-N, t}$ describes the data window of size N between $t - N$ and t, and X_{t+1} is the future probability at time $t + 1$.

$$Loss(\theta) = \ln(\cosh((P_\theta(X_{t-N, t}) - X_{t-1})))$$
(1)

It's a very simple strategy: with the given past tendency, predicting next year's word appearance tendency. We applied Adam optimizer [13] for stochastic weight readjustment.

We used TensorFlow 2.9.1 for implementation. And a MacBook Air with Apple Silicon(M1) was used for training.
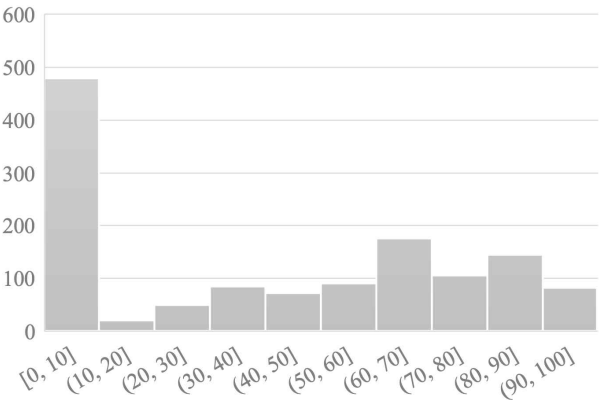
8. Normalization of Regression Results

<표 2> 가중합 정보
<Table 2> Weighted Sum Information

Window Size	Weight	Mean. Output.
3	0.5	13.14
5	0.4	15.07
7	0.3	15.86
10	0.2	15.32
13	0.1	14.07

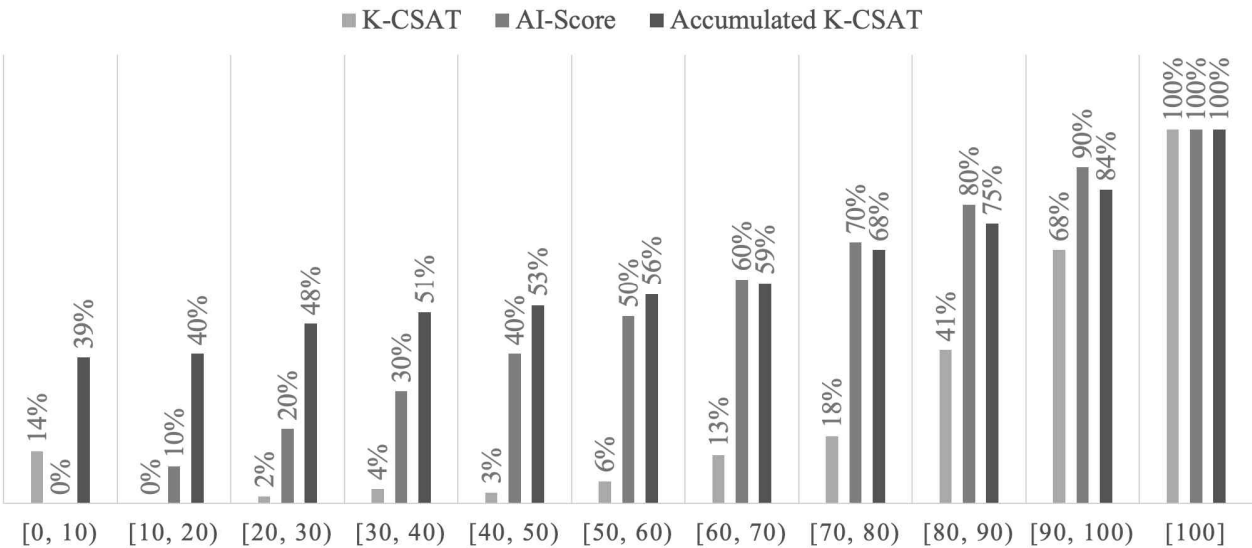[그림 4] AI-Score 분포
[Fig. 4] Distribution of AI-Scores

We trained the same model with 5 different-sized windows. After training the model with K-CSAT between 2003~2022, the trained model predicted frequency at 2023 exam, which was going to be held on November 17, 2022.

The predicted values for 5 different windows are described as 5 separated vectors with 1,300 values. Each value from different window was merged into single value by simple weighted sum. Window sizes used for the experiment and the weights are described on Table 2.

As larger window values implies long-ranged timeseries tendencies while smaller windows more concentrate on recent exams, we provided smaller weights for larger windows and bigger weights for smaller windows. We call the processed weighted sum as AI-Score.

[그림 5] AI가 예측한 점수의 구간별 정확도 분포
[Fig. 5] Illustration of the Accuracy of AI-predicted Scores

IV. Result

1. AI Prediction Values

Processed weighted sum of AI predicted values is available at ResearchGate (Supplement Data 6) and the distribution is described on Fig 4. The Model predicted 479 words will not appear on 2023 K-CSAT Exam.

2. Prediction of 2023 K-CSAT Exam

17, 2022, we obtained and processed the English exam PDF file into processed tokens. Total 1,558 tokens including some misprocessed symbols or conjugations of verbs were extracted. After removing the words which we have screened to remove artifact, 1,120 meaningful words remained. The list of words are available online (Supplement Data 7).

<표 3> 예측 결과
<Table 3> Prediction Results

	Proposed Method	Previous Work [2]
Number of Words in interest*	1,300	3,208
Accuracy**	52.5%	12%
Intersection Rate***	68.2%	34.4%

* The number of words in interest

** (# of True Positive) / (# of words in interest)

*** (# of True Positive) / (# of words from 2023 K-CSAT)

After 2023's K-CSAT exam was held on November The prediction results of proposed methods and previous works are summarized on Table 3. Our method showed 4.35 times higher accuacy compared to Word Master.

Also, the rate of intersection between the set of words from 2023 K-CSAT and the set of the words in interest of the proposed method is 1.98 times higher than Word Master.

Although Copyright law is limited on research purpose so providing word lists from those two book is legally possible, we would not provide it, because the lists are the most important intellectual property of their books. As many people are involved in their publication business, we do not want to disrupt their economic activities. Therefore, if you want to access the list of those two books, please reference the printed books; ISBN codes are available on the reference section.

4. AI Prediction Accuracy

After 2023's K-CSAT exam was held on, we compared the words with our AI-Scores. The results are described on Fig 5. The values of K-CSAT explains the appearance rate of words between each AI-Score segments. And the values of Accumulated K-CSAT is the rate of total words, in each segments and higher segment: in short, it's an accumulated rate calculated from the top.

V. Discussion

AI-Score works more similar to accumulated K-CSAT distribution rather than K-CSAT result itself. Also, AI-Score worked much better at high-score segments rather than low-score segments. The result showed that ground truth probability of English vocabulary appearance was much higher than AI predicted value.

Therefore, in other words, we assume that any words which AI predicted to be likely to appear on the next exam would appear more often actually. Also, with the comparison result described on Table 3, we think that applying human experts' know-how on data driven analysis model would perform much better.

One weak point of our approach was the size of data. As the data appearance pattern between

K-CSAT and mock test was too different, we could not directly train our model on mock test to predict K-CSAT patterns. Therefore, if more data were available, much more dramatic result would be expected.

VI. Conclusion

We have provided a set of words with high correlation with K-CSAT. We provided a preprocessing modules for K-CSAT exam tokenization, and for word appearance tendency analysis. After 2 human experts' screen, 1,300 words remained. Those words set showed much more accuracy rate than previous methods.

Also, we suggested an LSTM based AI model which learns the pattern of vocabulary appearance from K-CSAT exams. It showed higher accuracy on high-probability sections, and lower-probability words tend to appear much more than we expected.

Acknowledgement

We would like to express our great gratitude to Jehun Ryu for OCR-recording of word sheets from *Word Master*.

Supplement Data

[1] https://github.com/needleworm/bigdata_voca/tree/main/pdf%20parser
[2] https://github.com/needleworm/bigdata_voca/blob/main/2023%20sunung%20analysis/extract_sentences_from_sunung.py
[3] https://github.com/needleworm/bigdata_voca/blob/main/bigdata%20analysis/train.py
[4] https://github.com/needleworm/bigdata_voca/blob/main/2023%20sunung%20analysis/extract_sentences_from_sunung.py
[5] https://www.researchgate.net/publication/3654911 64_sunung_english_analysis
[6] https://www.researchgate.net/publication/365491618_ai_score_of_words
[7] https://www.researchgate.net/publication/365543324_2023_K-CSAT_words

References

[1] "대학수학능력시험 연도별 접수현황", Korea Institute of Curriculum & Evaluation. 2022. Web Report. 16 Nov. 2022.
[2] Hyungjung Kim, 워드마스터 수능 2000, 5th edition, Seoul: Etoos Edu Inc., 2021. Print. ISBN: 979-11-6123-728-2. 2000.
[3] Seonsik Kyeong, 경선식 수능영단어 Vol.1 기본, 2022 edition, Seoul: Kyong Seon Sik Edu. 2022. Print. ISBN: 979-11-8990-217-7, 2022.
[4] Hyerang Lim. "Predicting the Item Difficulty of a Simulated CSAT English Test Based on Corpus Analysis." 영어평가 16(1), 59-78, 2021.
[5] Wang, Yichuan, et al. "A Heuristic SAT Problems Solving Method based on LSTM Network.", 2021 12th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP). IEEE, 2021.
[6] Zhang, Chenhao, et al. "Towards Better Generalization for Neural Network-Based SAT Solvers." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2022.
[7] Serban, IV, García-Durán, A., Gulcehre, C, Ahn, S, Chandar, S, Courville, A and Bengio, Y, Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. ACL, 2016.
[8] Xu, Xing, et al. "Radial graph convolutional network for visual question generation." IEEE transactions on neural networks and learning systems 32(4), 2020: 1654-1667.
[9] Tang, Duyu, et al. "Question answering and question generation as dual tasks." arXiv preprint arXiv:1706.02027, 2017.
[10] Mikolov, Tomas, et al. "Efficient estimation of

word representations in vector space." arXiv preprint arXiv:1301.3781, 2013.

[11] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9(8), 1735-1780, 1997.

[12] Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." Multimedia Tools and Applications, 1-32, 2022.

[13] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.