International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 1
ISSN: 2088-5334

# Doc2Vec based Question and Answer Search System

HeeSeok Cho[a,1], Yong Kim[b,2]

[a] 551, Eonju-ro, Gangnam-gu, Seoul, Republic of Korea
[b] Department of e-learning, Graduate School, Korea National Open University, 86 Daehak-ro, Seoul, 03087, Republic of Korea
Corresponding author: [1]hscho@tekville.com, [2]dragonknou@knou.ac.kr

***Abstract*— E-learning interaction acts as a positive factor, such as improving learning commitment and learning effect and reducing the dropout rate. As an important function of e-learning interaction, if a learner queries a content that is difficult to understand during learning, a question-and-answer bulletin board that responds to the question is provided by a professor. In the way that the instructor directly answers the learner's questions, real-time feedback is difficult, and the instructor's fatigue increases. The purpose of this study is to achieve the goal of reducing answering time and reducing answering costs by developing a question-and-answer search system that automatically searches for and provides answers to questions created by learners during learning. To this end, this study designed and implemented a question-and-answer search system that provides the most similar query answers to learners by analyzing questions and answers based on Doc2Vec, one of the word embedding technologies, which is a natural language processing technology. By applying the results of this study to the question-and-answer system, it is expected that the learning effect can be enhanced by providing an immediate answer to the learner's question. In addition, organizations that pay response fees through the national budget, such as the Korea Educational Broadcasting Corporation, will be able to focus more on investments such as improving content quality through budget reduction.**

***Keywords*— eLearning; LMS; word2vec; doc2vec; question and answer search.**

*Manuscript received 24 Jul. 2020; revised 28 Aug. 2020; accepted 10 Dec. 2020. Date of publication 28 Feb. 2021.*


## I. INTRODUCTION

According to the 2018 Investigation on status of e-learning Industry, e-learners responded with a decrease in concentration when learning (30.4%) and experience of inconveniences with questions and responses (23.9%) as major inconveniences and problems with e-learning [1]. The EBS High school online tutoring system presented by Korea's Educational Broadcasting System provides students with question and answer service where they can ask questions. Delay in response time in the question, and answer systems is a major inconvenience to the learners. And 48.84% of the survey respondents stated that they require a "two-way interactive communication service enabling immediate feedbacks when asked questions about EBS textbooks and online tutoring videos" [2].

A remote study on adult learners also found that 'question and answer system about curriculum' as a type of online interaction activity is the one that learners consider most important, along with 'learning activity management/encouragement' and 'assignment castigation' [3]. This study is the design and implementation of a system that uses the Doc2Vec algorithm of word-embedding technology, one of the natural language processing techniques, to search the database for questions and answers similar to those proposed by learners and provide them with live feedback. Question and answer search system based on Doc2Vec is a system that can perform non-guidance learning with Doc2Vec algorithm after pre-processing question and answer data stored in the database and search for accurate questions and answers to perform as an immediate response to the queries. This study aims to improve the issues where we are unable to provide 24-hour feedback from our problem and answer service. Through this system, learners can get answers to questions in real-time, and service providers can improve learners' discomfort.

## II. MATERIALS AND METHOD

### A. Word Embedding

Word Embedding is a technique that expresses words in the form of vectors that implies order and meaning and is highly utilized in natural language parsing. Word embedding begins with the premise that words with the same context have close

distance within the vector space. Word2Vec is a representative algorithm for word-imbedding and is one of the most useful techniques in natural language parsing [4]-[6].

Doc2Vec is an algorithm that extends word embedding Word2Vec, expressing the document itself as a fixed-sized vector containing the meaning, using the words included in the document. Doc2Vec is used in the analysis of text documents such as articles, web documents, comments, and content metadata, and the Gensim Doc2Vec algorithm is used a lot in Python development environment [7]-[9].

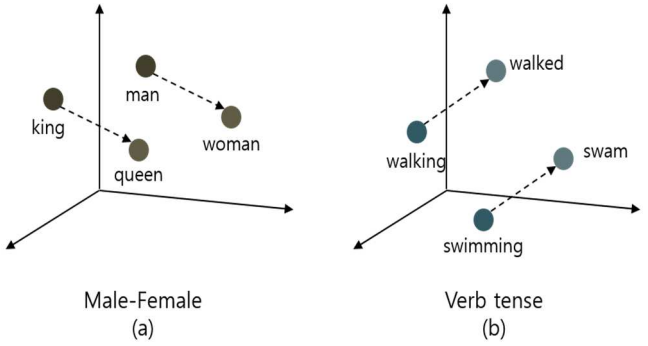


Fig. 1 Word embedding

Fig. 1 (a) above indicates a semantic inference that when you remove "man" and substitute it with "woman" within the "king" vector, it can produce "queen" as the result. Fig. 1 (b) shows an example of grammatical reasoning, uses the fact that the distance between "walking" and "walked" may be equal to the distance between "swimming" and "swam," to present the possibility of studying the meaning of a word according to the relative context within the vector space.

### *B. Korean Alphabets' Morphological Analysis*

The document-based language analysis model is used after the tokenization process of breaking documents into words. In the case of Korean documents, they will be tokenized through a Korean morphology analyzer to extract words and select only the parts necessary for natural language parsing by obtaining information of each word. Fig. 2 is the result of an analysis through a morphology analyser, which allows data that has completed a morphological analysis to filter specific components to increase the quality of the Doc2Vec model [10]- [12].

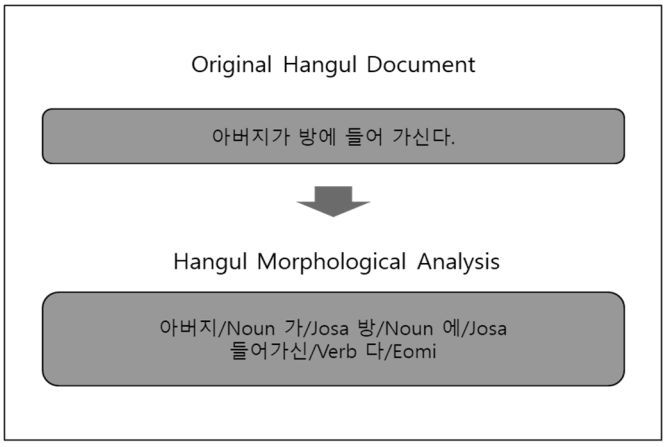


Fig. 2 Examples of hangul morphological analysis

## III. Result and Discussion

### *A. Process of the Text*

To achieve the objective of this study, the study shall proceed with the following process: procuring of the data, system design, development, verification. The 58,279 data required for the study was accumulated in the latest three years and was collected from Korea's Educational Broadcasting System's Q&A board about EBS "Social Studies: Civics" Highschool e-learning service. The system was designed with Use Case Diagram, Program and its' Deduction of Skills, Sequence Diagram.

On the developmental stage, data collector, data pre-processor, machine learning operator, question and answer search engine was materialized based on python. At the verification stage, primarily, the question and answer composition and the quality of machine learning model according to each pre-processing method were compared, and ultimately, the accuracy of the machine learning model was measured based on the 30 Q&A datasets. The development environment of the study is shown in Table I below.

TABLE I
Development Environment of the Study

| Category | Content of the Study | Name of Software | Version |
|---|---|---|---|
| Design | UML DIAGRAM-based design | STAR UML | 3.1.0 |
| Database | Saving of Q&A data | MariaDB | 10.1.38 |
| Development language | System Implementation | Python | 3.6.8 |
| Development Tools | Python Package including Science and Mathematics Python | Anaconda | 4.6.14 |
| | Integrate Developmental Environment | eclipse | 2019-03 (4.11.0) |
| Data Analysis tools | Python library for topic modelling, document indexing and similarity retrieval with large corpora | Genism | 3.8.1 |
| | Visualization with Python | Matplotlib | 3.2.1 |

### *B. System Design*

The question and answer search system provides users with the most similar search result from what was written by the user and what was saved in the database, as shown in Fig. 3 Use Case Diagram. The system aims to minimize and to make convenient, the time-delayed for the learners to receive the answer to their inquiries by searching the similar one using Doc2Vec, taking account of the fact that there are numerous similar questions and answers on e-learning.

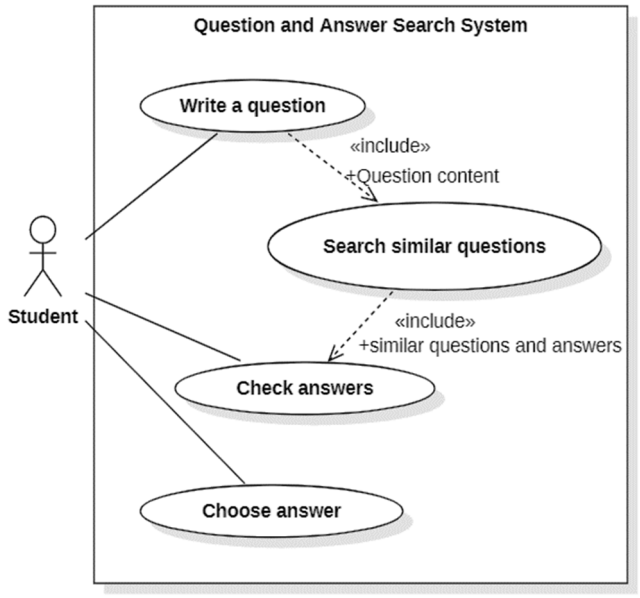


Fig. 3 Use case diagram of Q&A search system

Fig. 4 explains the data processing flow of the question and answer search system. The system is generally divided into 1. Creation of a Doc2Vec model constructing the dataset with distinct ID documents through data pre-process saved in the total database, and 2. The searching stage where when a learner puts in a question, Doc2Vec model, created searches for a similar question and answers.

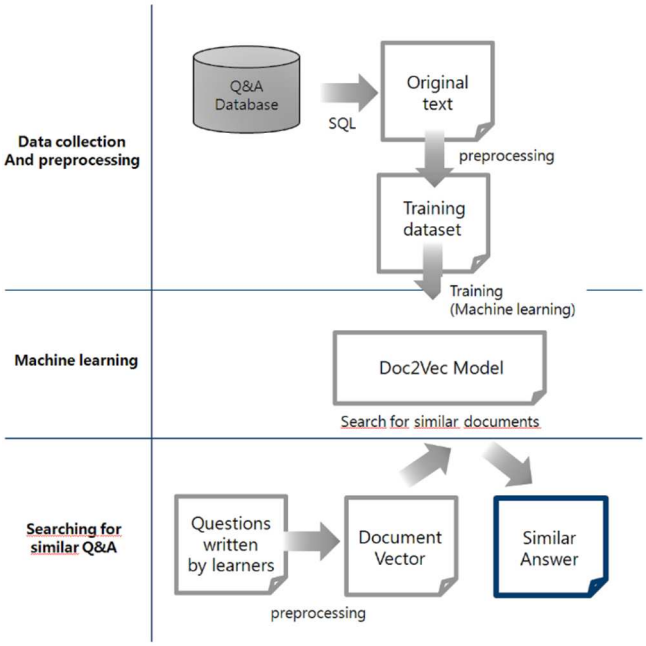


Fig. 4 Data flow of Q&A search system

The system was configured to perform analysis of Korean Alphabets' morpheme, filtering of words, and processing of non-verbal terms in the generation of machine learning models and question and answer searching system. System implementation takes a long time to pre-process because it takes time to go through the entire question and answer data stored in the database. Therefore, the system should be designed and implemented by storing pre-processed results in separate files and subsequently pre-process and merge newly stored Q&A data only.

Table II defines the function of similarity search program based on Doc2Vec model, which pre-processes what was put in by a learner and searches for a similar question and answer from the database. Doc2Vec based question and answer search engine is composed of data collector, data pre-processor, machine learning activator, question and answer searcher.

TABLE II
FUNCTIONING COMPOSITION OF Q&A SEARCH SYSTEM

| Name of the Program | Function | Explanation |
|---|---|---|
| Data Collector | SQL | Collection of Q&A data from the database using SQL Statement and saving the data with the question's distinct key in text format |
| | Morpheme Analysis | Morpheme analysis of Korean morpheme analyzer. Saving the dataset of the document with the distinct key for Doc2Vec machine learning |
| Data Pre-processor | Stop word removal | Processing Korean stop words with user defined stop words |
| | Saving of Dataset | Saving and conversion of the dataset ready for Doc2Vec machine learning |
| Machine Learning Generator | Machine Learning | Processing of Doc2Vec machine learning using word lists for each document IDs |
| Q&A Search Engine | Pre-processing of questions | Pre-processing of morpheme analysis as of machine learning generation process |
| | Search | Search for Q&A that is similar to the one presented by the learner |

In general, morpheme analysis is carried out at the pre-processing stage. However, it takes much time to morpheme large-sized text. In this study, it is important to see how the search quality varies for different text processing methods. Morphological analysis, which is applied equally to all experiments, saves the results in advance to shorten the experiment time.

### C. Data Collector

The key role of the data collector is to execute Q&A data stored in the database by executing SQL statements. Q&A data should be searchable by grade and subject, so it should be designed in consideration of this. Data collector saves the results from the morpheme analysis as files, and the results are what has been processed through SQL statements from the question and answer database. The sequence diagram of the data collector is as shown in Fig. 5.

The morpheme analysis of this study was performed with Twitter morpheme analyzer using Python. Doc2Vec needs to be composed of dataset, also known as the set of documents with distinct key, in order to be applied to machine learning. Since the quality of Doc2Vec model can vary according to each dataset compositions, the system have been implemented

to be able to compare the different quality according to different compositions of title of the inquiry, content of the inquiry, content of the answer within the question and answer database.

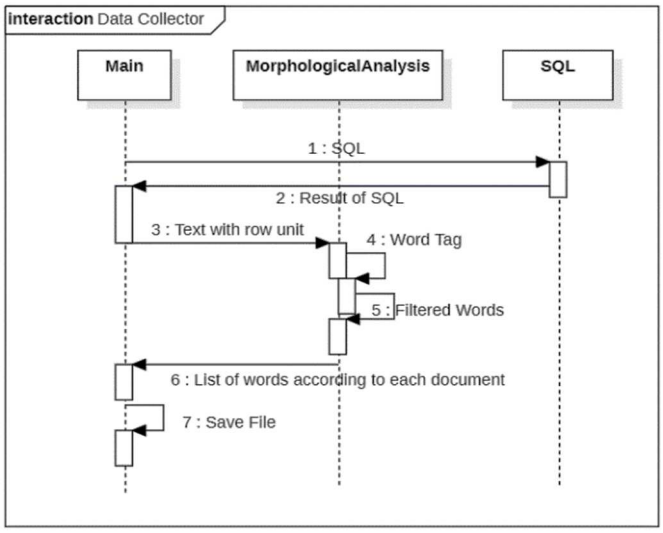


Fig. 5 Sequence diagram of data collector

### D. Data Pre-processor

The data pre-processor performs its role in deleting stop words and/or words that have been highly frequently used in majority of the documents to increase the quality of machine learning model. The sequence diagram of the data pre-processor is as show in Fig. 6.

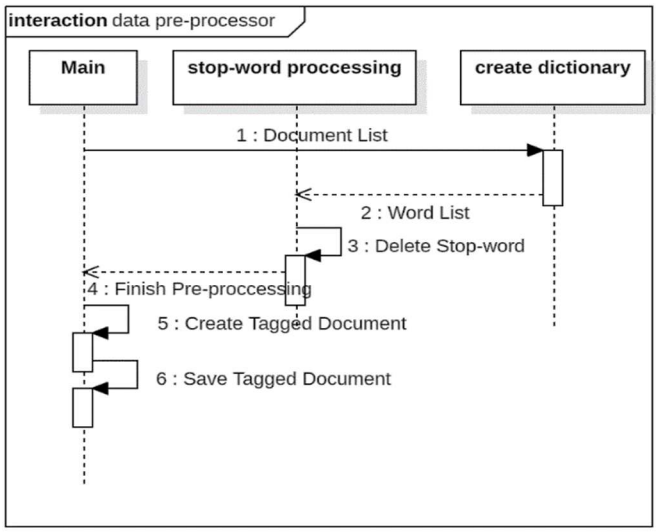


Fig. 6 Sequence diagram of data pre-processor

For this study, Korean stop words have been removed according to what was provided by Ranks NL (https://www.ranks.nl/stopwords/korean). Further, high frequency of identical words lowers the quality. Therefore, they must be deleted. In order to implement this factor into the system, a word dictionary was made from distinct ID and documents of word lists from the data collector, and based on that same word dictionary, removal of words that have been used more than 10% have been processed by calculating the word frequency of each document and the document frequency including those words frequently used[20].

### E. Machine Learning Operator

The machine learning operator plays its role in saving the results of pre-processed datasets and what have also been processed through machine learning using Doc2Vec algorithm, as the model file. The Sequence diagram of machine learning operator is as shown in Fig. 7.

Doc2vec is basically a machine learning algorithm extended to documents based on word2vec. Therefore, the doc2vec model supports word-related API of word2vec. To test the basic model, we searched for similar words. Based on the generated machine learning model, similar words for the word 'genetics' were searched for words with semantic similarities such as 'physical', 'science' and 'biology' as shown in Table III.

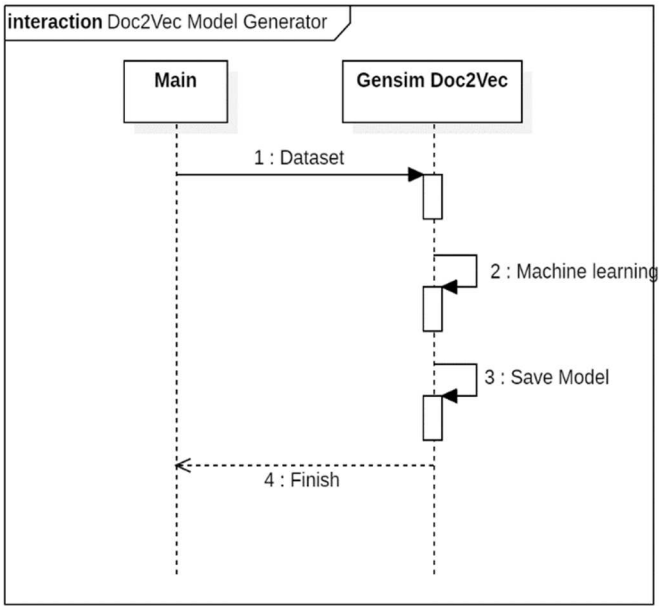


Fig. 7 Sequence diagram of machine learning operator

TABLE III
RESULT OF SYNONYMS OF THE WORD "GENETICS"

| **Searched Word** | **"Genetics"** |
|---|---|
| **Proximity of Result of Word Similarity (0~1)** | [('Physical', 0.6417882442474365), ('Science', 0.6417572498321533), ('Biology', 0.640328049659729), ('Duplication', 0.6352273225784302), ('Mechanism', 0.6346608996391296), ('Species', 0.6322759985923767), ('Arts', 0.6234589219093323), ('Embryo', 0.622969388961792), ('Scientific Technology', 0.6219070553779602), ('Biological Experiment', 0.621895968914032)] |

### F. Q&A Search System

The question and answer search engine uses Doc2Vec model intended from Doc2Vec model generator to find similar question and answer that the learner presents. Sequence diagram of question and answer search engine is as shown in Fig. 8.

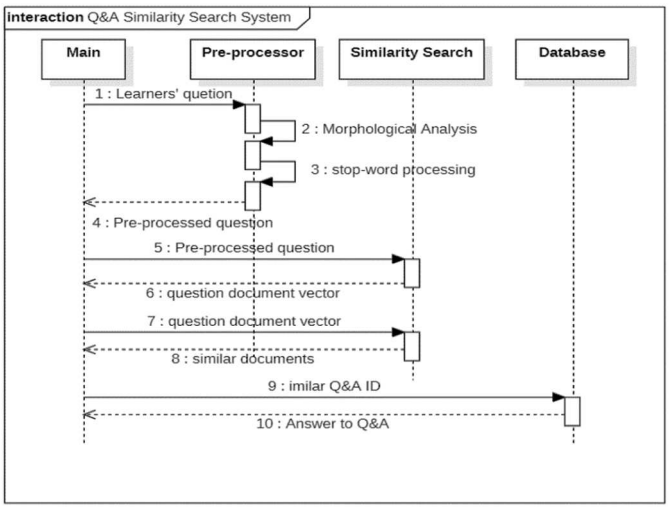


Fig. 8 Sequence diagram of Q&A search engine

Once the learners have filled out and saved the query, the stored query is converted to a document vector so that it can be searched in the Doc2Vec model after the pre-processing process. The Doc2Vec model provides an equation to convert documents into vectors, so it is simple to convert documents into vectors. The learners' queries converted to vectors are then presented to learners using a similar vector search equation in the Doc2Vec model to search for questions and answers with the most similar content stored in the database.

### *G. Result of the Study and Quality Verification*

The Doc2Vec model for retrieving document similarity in natural language processing depends on how the corpus is constructed for machine learning and pre-processing. This study has put the accuracy of question and answer search engine to test by composing title of the question, the content of the question, the content of the answer differently and also by differentiating the deletion of stop words and selection of parts of speech at the pre-processing stage, in order to find out the quality of Doc2Vec based search engine and correlation of methods of pre-processing.

Result of the number of similar questions and answers included within the top 3 similarity score, similarity according to each pre-processing stage, and the composition of question and answer document is shown as Table 4 and Table 5 below. Table IV shows the results of deleting the terminology and pre-processing it. In the table, the label N.S. is the number of very similar contents in TOP3.

TABLE IV
QUALITY OF SEARCH WHEN STOP WORDS ARE DELETED

| Pre-processing Category | Noun | | Noun + Verb | | All morphemes | |
|---|---|---|---|---|---|---|
| | Highest Score | N.S. | Highest Score | N.S. | Highest Score | N.S. |
| **Full Text** | 76.89% | 2 | 80.66% | 2 | 79.54% | 2 |
| **Title+ Question** | 78.00% | 3 | 79.67% | 3 | 86.95% | 2 |
| **Title+ Question + Answer** | 82.22% | 3 | 83.85% | 3 | 76.70% | 3 |
| **Answer** | 77.11% | 1 | 67.89% | 1 | 78.29% | 1 |
| **Question + Answer** | 69.45% | 2 | 75.00% | 2 | 75.09% | 1 |

Table V shows the results of pre-processing without deleting the stop words.

TABLE V
QUALITY OF SEARCH WHEN STOP WORDS ARE NOT DELETED

| Pre-processing Category | Noun | | Noun + Verb | | All morphemes | |
|---|---|---|---|---|---|---|
| | Highest Score | N.S. | Highest Score | N.S. | Highest Score | N.S. |
| **Full Text** | 85.25% | 2 | 81.79% | 2 | 86.28% | 2 |
| **Title+ Question** | 86.92% | 2 | 85.22% | 2 | 85.28% | 3 |
| **Title+ Question + Answer** | 83.67% | 3 | 83.95% | 3 | 83.23% | 3 |
| **Answer** | 72.97% | 1 | 78.46% | 1 | 81.74% | 1 |
| **Question + Answer** | 67.63% | 3 | 79.50% | 2 | 77.10% | 1 |

As shown in Table VI within the Doc2Vec document dataset, the highest possible quality could be obtained, combining all three: title of the question, the content of the question, and content of the answer. Also, when stop words were deleted, and the usage of nouns and verbs in morpheme analysis was proceeded.

TABLE VI
RESULT ANALYSIS ACCORDING TO DOC2VEC DOCUMENT DATASET COMPOSITION AND PRE-PROCESSING METHODS

| Experiment Method | Highest Quality |
|---|---|
| **Q&A Document Composition** | All title of inquiry, content of inquiry, content of answer |
| **Choice of Parts of Speech** | Parallel use of noun and verb |
| **Stop Word Processing** | Stop word removal |

In Doc2Vec algorithm, with quality of pre-processing as well as the dataset composition, parameter for machine learning was one of the categories improving the quality. Parameter used in the study is as shown in Table Ⅶ.

TABLE VII
PARAMETER OF DOC2VEC MACHINE LEARNING USED IN THE STUDY

| Name of Parameter | Explanation | Value |
|---|---|---|
| **window** | Set considering adjoining words in front and back | 3 |
| **min_count** | Minimum frequency of words used in the data | 10 |
| **max_epochs** | Number of optimizations | 100 |
| **vec_size** | Size of embedding vector | 100 |
| **workers** | Number of CPU core for parallel arithmetic | 8 |

To find out the basic characteristics of the generated Doc2Vec model, I searched similar words with a few words. As a result, as shown in Table VIII, words that are semantically similar are searched [13], [14]. In order to identify the quality of machine learning model, dataset used for the quality assessment must be prepared beforehand. In this study, experiment dataset was composed by separating 30 questions from the total question and answer data, and this data was exempt from learning dataset that generated the final Doc2Vec model as the study proceeded [15].

TABLE VIII
SIMILAR WORD SEARCH RESULTS

| word | Similar word search results |
|---|---|
| **Disobe-dience** | World, disobedience, minority, disobedience, resistance, war, ownership, death penalty, argument, calling |
| **Philoso-pher shun** | Private Interest, Durure, Poverty, Marx, Orthopedic, Puda, Yoonsa, Death, Officials, Farmers |
| **democracy** | Democracy, Private, Common, Upper Class, Absolute, World, Justice, Hangout, Legal, USA |
| **Marx** | Spinoza, Jeong Yak-yong, Controversy, Nivoir, Jaspers, Sunja, Plato, Non-interference, Nationalism, Mencius |
| **Hobbs** | Reverse discrimination, SeonSeonSeol, Beopbo, First, Justice, Resistance, Mani, Substitution, SeongakSeol, Manjangil |
| **female** | Rescue, counterpart, incline, have, male, manual labor, servant, ethnicity, husband, market failure |

Table IX is the result of the search quality of the question and answer search system implemented with the foundation of Doc2Vec. The result of Q&A Search evaluation with 30 pre-documented questions and answer queries, as shown in Table IX, presented that similar question and answer search average ranking 1st resulted in 67.9%, with a 60% probability of a very similar Q&A that ranked 1st place. The probability of a very similar question and answer searched within the top three of the search results was 70%.

TABLE IX
QUALITY OF DOC2VEC-BASED Q&A SEARCH SYSTEM

| Criteria | Average of Similarity of first results | Probability of very similar Q&A searched as first result | Probability of very similar Q&A searched within top three search results |
|---|---|---|---|
| **Score (Average)** | 67.9% | 60% | 70% |

## IV. CONCLUSION

Q&A in e-learning is an important interactive tool that can motivate learners and enhance their learning effectiveness. However, in the current process where the educator identifies a learner's question and writes a response, limitations in immediate feedback do occur [16],[17]. This study aims to provide learners with immediate feedbacks on their queries based on word embedding algorithm technology: Doc2Vec, processing unsupervised machine learning, implementing questions and answer search system with database of previous queries.

A question and answer search based on the generated Doc2Vec model produced a 60% probability of providing a very similar answer with the first search result 70% for the top three search results. What was only available with expensive search engines, can now be simply applied with the results of this study in document similarity searches and if reflected in actual services, there is a 70% chance that students will be supplied with adequate answers to their queries.

If you want to provide real-time answers to the current question-and-answer system, you can use Doc2Vec algorithm to provide services with sufficiently good performance and quality. The use of big data and artificial intelligence has become more active in various fields. Natural language parsing is available in analyzing various unstructured data required by e-learning [18]-[20]. The study anticipates using the results in analysing various unstructured data in e-learning field, which will further develop into an artificial intelligence question and answer service based on knowledge. Recently, the importance of non-face-to-face education is increasing worldwide due to covid-19. It is expected that this study will contribute to enhancing the learning satisfaction and enhancing the learning effect by resolving the curiosity of learners in real-time in increasing face-to-face education.

## REFERENCES


[1] National IT Industry Promotion Agency(NIPA), *Investigation on status of e-learning Industry*, Ministry of Trade and Industry(MTI), 2018.

[2] T.H. Kang, *Case Studies on EBS CSAT*, Educational Broadcasting System, 2016.

[3] LAFLEN, Angela, SMITH, and Michelle,"Responding to student writing online: Tracking student interactions with instructor feedback in a Learning Management System," Assessing Writing, vol 31,pp. 39-52. 2017.

[4] W. Chang, Z. Xu, S. Zhou, Shenghan,and W. Cao, "Research on detection methods based on Doc2vec abnormal comments," Future generations computer systems,vol.86, pp. 656-662, 2018

[5] D.H. Kim, D.S. Seo,S. Y. Cho,and P. S. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," Information sciences,vol. 477 pp. 15-29. 2019.

[6] B. Pan, C. C. Yu, O. C. Zhang, S. X. Xu, and S. Cao, "The Improved Model for word2vec Based on Part of Speech and Word Order," ACTA ELECTRONICA SINICA vol. 46, pp 1976-1982, 2018.

[7] V. A.Nrusimha,W. Hannah,H. J. David,K.Matcheri, and T.J. Blake, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape," Canadian journal of psychiatry, vol. 64, pp. 456-464, 2019.

[8] M. Tomas, "Efficient estimation of word representations in vector space," arXiv preprint arXiv, 2013.

[9] M.Mudasir, J.Rafiya, and S. Muzaffar, "Text document summarization using word embedding," Expert systems with applications, vol. 143, 2020.

[10] (2017) Python Korean NLP website. [online]. Available: https://replet.tistory.com/58

[11] (2015) KoNLPy website.[online]. Available: https://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/#comparison-between-pos-tagging-classes

[12] J.Yaser, A. A. Mahmoud, and B.Elhadj, "Advanced Arabic Natural Language Processing (ANLP) and its applications: Introduction to the special issue," Information processing & management,vol. 56, pp. 259-261, 2019.

[13] C.Jingqiang,Z. Hai, "Extractive summarization of documents with images based on multi-modal RNN," Future generations computer systems, vol. 99, pp. 186-196, 2019.

[14] Y. Cheng, Z. Ye, M.Wang, and Q. Zhang, "Document classification based on convolutional neural network and hierarchical attention network", Neural Network World, vol. 29, pp. 83-98, 2019.

[15] F. Yang,F.Lidan, "Ontology semantic integration based on convolutional neural network,"Neural Computing And Applications, vol. 31, pp. 8253-8266, 2019.

[16] I. Alsmadi,H. G.Keng, "Term weighting scheme for short-text classification: Twitter corpuses,"Neural Computing And Applications, vol. 31, pp. 3819-3831, 2019.

[17] S. N. Bhushan, A. Danti, "Classification of text documents based on score level fusion approach," Pattern recognition letters, vol. 94, pp. 118-126, 2017.

[18] M.Chowkwanyun, "Big Data, Large-Scale Text Analysis, and Public Health Research," American journal of public health, vol. 109, pp. S126-S127, 2019.

[19] Y. Q. Song, U.Shyam,P.Haoruo,M. Stephen, and R. Dan. "Toward any-language zero-shot topic classification of textual documents" Artificial intelligence,vol. 274, pp. 133-150, 2019.

[20] K. Hu, H. Wu, K. Qi, J. Yu , S. Yang, T. Yu , J. Zheng, andB. Liu, "A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model," Scientometrics, vol. 114, pp. 1031-1068, 2018.