J Inf Process Syst, Vol.18, No.2, pp.282~291, April 2022
https://doi.org/10.3745/JIPS.04.0239

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A BERT-Based Automatic Scoring Model of Korean Language Learners' Essay

Jung Hee Lee[1], Ji Su Park[2], and Jin Gon Shon[3,*]

Abstract

This research applies a pre-trained bidirectional encoder representations from transformers (BERT) hand-writing recognition model to predict foreign Korean-language learners' writing scores. A corpus of 586 answers to midterm and final exams written by foreign learners at the Intermediate 1 level was acquired and used for pre-training, resulting in consistent performance, even with small datasets. The test data were pre-processed and fine-tuned, and the results were calculated in the form of a score prediction. The difference between the prediction and actual score was then calculated. An accuracy of 95.8% was demonstrated, indicating that the prediction results were strong overall; hence, the tool is suitable for the automatic scoring of Korean written test answers, including grammatical errors, written by foreigners. These results are particularly meaningful in that the data included written language text produced by foreign learners, not native speakers.

Keywords

Automatic Writing Scoring, Bidirectional Encoder Representations from Transformers, Korean as a Foreign Language, Natural Language Processing

1. Introduction

The most important part of descriptive scoring is determining whether the scoring is consistent based on valid criteria [1]. Therefore, if these scoring criteria and processes can be automated, scoring costs and the temporal burdens of large-scale evaluations could be greatly reduced. Additionally, learners could improve their writing skills faster by receiving immediate feedback on their written inputs. To this end, it is necessary to determine whether automatic scoring is possible for Korean learners' descriptive answers containing various errors of form, syntax, and grammar. Research related to automatic scoring capabilities for descriptive Korean-language text-exam responses has developed alongside corresponding natural language processing (NLP)-related technologies [2-4]. A language model performs the allocation and probability determination of word sequences and sentences in NLP [5]. Machine- and deep-learning technologies, such as the embeddings-from-language model, the generative pre-training model, and the bidirectional encoder representations from transformers (BERT), have been developed for NLP tasks [6,7]. Unfortunately, their application to automatic scoring systems is lacking [8].

In this paper, a pre-trained BERT model is used to score foreign Korean-language students' descriptive

Manuscript received September 13, 2021; first revision November 17, 2021; accepted November 29, 2021.

*Corresponding Author: Jin Gon Shon (jgshon@knou.ac.kr)
[1] Dept. of Korean Language Education as a Second Language, Kyung Hee University, Seoul, Korea (iiekor@khu.ac.kr)
[2] Dept. of Computer Science and Engineering, Jeonju University, Jeonju, Korea (jisupark@jj.ac.kr)
[3] Dept. of Computer Science, Korea National Open University, Seoul, Korea (jgshon@knou.ac.kr)

written test answers, and the results are compared to other language model techniques. BERT's pre-trained model can provide consistent results when trained on a small number of data [6]. Fig. 1 displays the high-level BERT model.

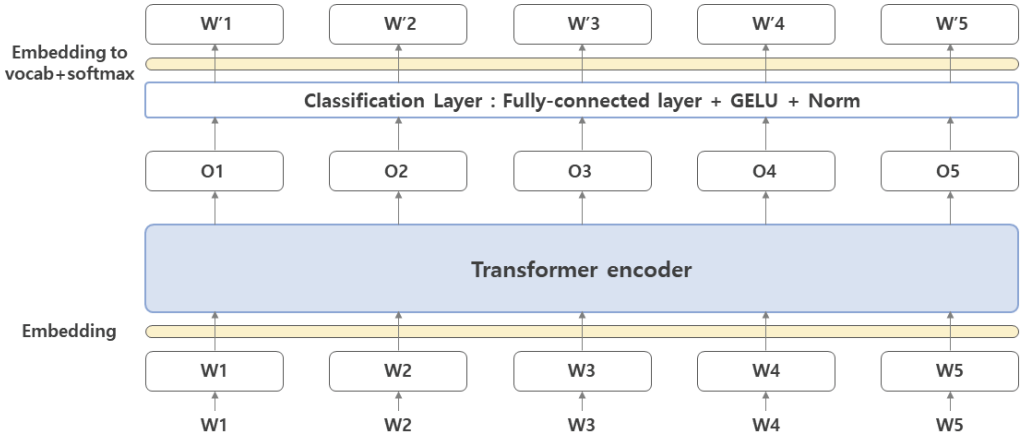

Fig. 1. BERT model.

2. Characteristics of Descriptive Answers of Korean Learners

2.1 Intermediate Korean Learner's Language Characteristics

There are 200 training hours for each grade (10 weeks × 20 hours) according to the International Korean Curriculum Standard for regular domestic institutions and the Korean Language Proficiency Test, governed by the National Institute of Korean Language. The data used in this paper correspond to Intermediate 1 learners, whose goals include being able to write abstract topics about familiar social life in a simple structure [9]. At the intermediate level, learners have accomplished 350–400 hours of Korean language learning and have acquired about 100 grammar knowledge items and 1,000 vocabulary words.

These learners can create complete sentences, but there are often errors in word order. Complexity increases when developing various sentence structures related to life experiences. Basic errors that often appear are related to the persistence of the grammar systems applied in their native language. Students have a strong tendency to ask teachers for detailed explanations to help them reconcile the new rules to their old rules. However, in terms of vocabulary learning, students are interested in various media sources for learning and immersion. Furthermore, learners who are not familiar with Chinese characters suddenly realize the need to learn Chinese characters. Many structural vocabulary errors have been catalogued, including errors of derivation, behavior, and instruction [10].

In this study, data are obtained from past processes of teaching Korean to foreign students, where the foreign students are learning Korean. These students must acquire and master morphological, syntactic, and pragmatic language rules for the first time. The focus of this research is not on analyzing the quality of prose. Instead, it is about analyzing language rules in the context of midterm and final exams. This type of writing consists of personal introductions and simple expository writings.

2.2 Test Question Detail

The test questions for intermediate-level learners are not based on proficiency evaluations but on the curricula of educational institutions. Most learners score more than 70 points. For example, a test question may direct the student to write more than 500 characters about a memorable travel account and provide categorical details, inclusively, of destination, transportation, purpose, sights, feelings, events, and reasons using complete sentences.

3. A BERT-Based Automatic Scoring Model

3.1 Model Design

A BERT-Based Automatic Scoring Model consists of three modules: data analysis, data preprocessing, and data learning (Fig. 2).

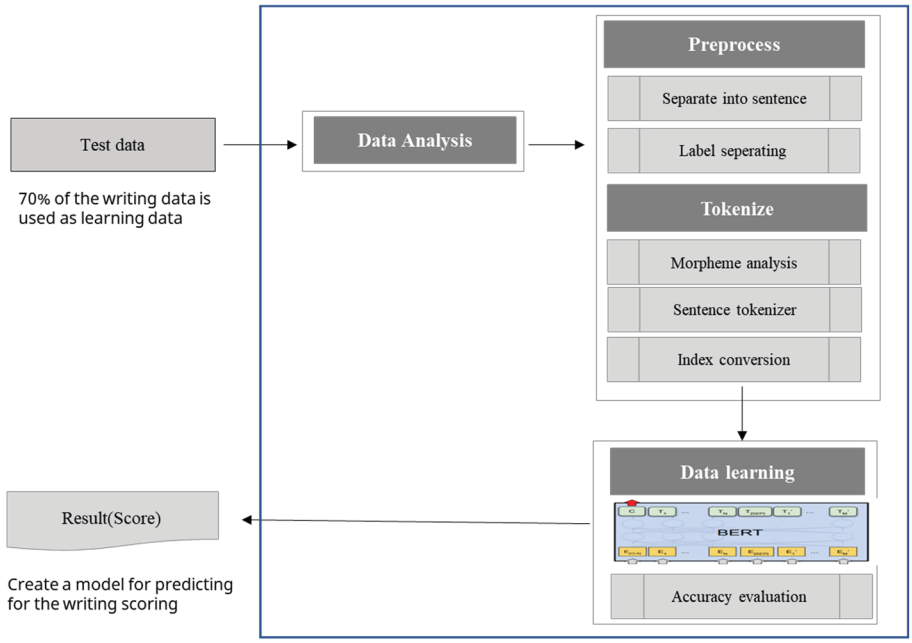

Fig. 2. A BERT-Based Automatic Scoring Model.

Because the actual tests were conducted as written exams in which answers were written on paper, the contents of the answers were entered as a text file that included all typos, spaces, non-grammatical items, and symbols as-written by learners. A tag of "unknown" was added if the writing could not be recognized. Items were analyzed based on the English special character ratio, the distribution of the answer length, and the graded answer data. Factors, such as special characters and numbers, are parts that interfere with learning. The lengths of the answers and the distribution of scores were correlated to the grades.

In the BERT's preprocessing module, sentence tokenization was required. A [CLS] indicator was attached to the beginning of the text token, and a [SEP] was added to the end. In succession, each token

was indexed, the length of each sentence was compared to the maximum length, and padding was applied for those not reaching the maximum. Then, an attention mask was generated.

BERT is a two-way language model in which context is considered based on a transformer model. BERT differs from other models in that it conducts unsupervised pre-learning in both directions. The language model provides the probability distribution for the order of words. Through this two-way learning, performance is improved during the fine-tuning stage using the BERT pre-learning model.

A BERT model requires powerful graphical processor units (GPUs), and ours was implemented using Google Colab's Tesla T4. Our study is based on the Hugging Face model based on Transformers for PyTorch and TensorFlow 2.0. The input value required an attention process to improve learning efficiency after fixing the length. In this study, the maximum length was set to 384 based on the average length and learning efficiency of the writing test answer. The pre-trained BERT model of Oh Yeon-taek used Google's SentencePiece to learn 32,000 vocabulary words using about 180 million sentences from Wikipedia, and news data it achieved 87.8% from KorQuAD [11,12].

3.2 Model Implementation

3.2.1 Data collection and input

The data contained information collected from 2013 to 2017 from the Korean language education institution at K. University in Seoul, Korea. The dataset included 586 answers about "my hobby" and "memorable travel," as obtained from Level-3 (Intermediate 1) writing answers for mid-term and final exams (Table 1). The general criteria of Test of Proficiency in Korean (TOPIK) Level 3 are described as follows [13]:

The individual has no problem doing normal, day-to-day activities and has the basic language skills to use various public facilities and maintain interpersonal relationships. He/she can understand and express him/herself regarding not only familiar and specific topics but also familiar social issues in paragraphs. He/she is also capable of distinguishing the basic features of colloquial language and literary language and can understand and use the two forms of language him/herself.

Table 1. Distribution of answer sheets by score

	Score											
	100	**90**	**80**	**70**	**60**	**50**	**40**	**30**	**20**	**10**	**0**	**sum**
Task 1 (my hobby)	5	47	76	62	82	32	30	15	8	12	2	371
Task 2 (memorable travel)	8	48	44	55	35	12	6	4	2	0	1	215

3.2.2 Data analysis

From the data, the distribution of English special characters, answer length distributions, and score data were analyzed. In the case of "my hobby" 4.85% of sentences included question marks, 100% included periods, 8.63% included uppercase alpha characters, and 54.72% included numbers. In the case of "memorable travel," the ratio of sentences that included question marks was 3.72%, the period was 100%, the sentence including capital letters was 10.23%, and the sentence including numbers was 80.47%.

In the case of uppercase characters, proper nouns, such as place names and anthroponyms, were written in English. Factors, such as special characters, numbers, and punctuation, slowed calculation speeds,

interfering with learning. The distribution of answer lengths was as follows. Among the “my hobby” answers, the maximum length was 736 words, and the minimum was 57 words. The average was 506.19 words, the standard deviation was 114, and the median was 523. Among the “memorable travel” answers, the maximum length was 833 words, and the minimum was 137 words. The average was 618 words, the standard deviation was 112 words, and the median was 630 words. The length of the answer to “memorable travel” was longer than that of “my hobby.” Descriptive statistics are shown in Table 2.

The distribution of answers by score group is as follows. The average score of “my hobby” was 64.3, and that of “memorable travel” was 72.5. The median value was 70.

As shown in Tables 2 and 3, because the null hypothesis cannot be rejected: the correlation between the length of the answer and the score is verified. The results are shown in Table 4. As a result of checking the correlation between the score and the length of the answer, the topic, “my hobby,” has a correlation coefficient between the length of the answer and the score of 0.477. The correlation coefficient between the length of the answer and the score is 0.495.

Table 2. Descriptive statistics of answers by length

Task	my hobby	memorable travel
Samples	371	215
Mean of length	503	618
Median	523	630
SD	113	112
Minimum	57	137
Maximum	736	833
Skewness	-1.03	-1.21
Shapiro-Wilk W	0.946	0.93
Shapiro-Wilk p	<0.001	<0.001

Table 3. Descriptive statistics of answers by score group

Task	my hobby	memorable travel
Samples	371	215
Mean of length	64.3	72.5
Median	70	70
SD	19.9	16.7
Minimum	0	0
Maximum	100	100
Skewness	-0.747	-0.967
Shapiro-Wilk W	0.935	0.916

Table 4. Correlation between the lengths of the answers

		my hobby		memorable travel	
		Length of answer	Score	Length of answer	Score
my hobby	Length of answer	-	-	-	-
	Score	0.477 (<0.001)	-	-	-
memorable travel	Length of answer	-0.029 (0.667)	0.034 (0.623)	-	-
	Score	-0.071 (0.302)	0.011 (0.868)	0.495 (<0.001)	-

Values are presented as Pearson’s R (p-value).

3.2.3 Data pre-training

The dataset contained 586 answers from midterm and final exams of foreign Korean-language learners at Level 3 (Intermediate 1) Korean proficiency. The pre-training was carried out as follows. First, the writing test data of third-grade learners were transcribed into a text file. Second, personal information and topics were deleted using the Notepad++ program as a pre-processing process. Third, all personal data, carriage returns, punctuation marks, and special characters were removed. Then, single topics were integrated into single lines to facilitate text processing. Commands and scores were tagged at the ends of the lines to enable automatic processing.

Data processing was conducted based on the Hugging Face model. The use of GPUs is essential to BERT model success. There are several GPU methods, such as utilizing cloud services (e.g., Google Korab) or implementing a local environment. In this paper, Google Colab's Tesla T4 was used.

Data were prepared following transformer installation to load training and testing datasets. The data were divided into 70% versus 30% portions: one for training and the other for testing, respectively. During the verification stage, 10% of the training data were used. The next step was conversion according to the input format of BERT, and the Korean Sentence Splitter library was used to separate the converted content into sentence units. Then, the score labels were separated. The tokenizer step was next, which tokenized the content using BERT's tokenizer. This process converts tokenized sentences into numeric indices, calculates the maximum sequence length of the input token, converts each token into a numeric index, matches the sentence to the maximum sequence length, and fills the deficiencies with padded zeroes.

The next step initializes the attention mask, which sets the attention mask to one without padding and zero with padding. Padding is not performed in the BERT attention layer so that the speed can be improved. After separating the data into training and verification sets, the attention mask is also separated into training and verification. Then, the data are converted into a PyTorch tensor. Then, the batch size is set, and it is entered using PyTorch's DataLoader, where it is tied to the mask and labeled to set the data. During learning, data are imported based on the batch size, and to preprocess extracted review sentences, they are converted according to the input format of BERT; then labels are extracted.

3.2.4 Data learning

A bidirectional language model was created using BERT for data learning (Fig. 3). When dealing with text, this kind of model is advantageous because the problem of word ambiguity can be solved, even if the same word has different embedding vector values based on sentence form and location. The classification model leveraged 11 layers and an attention method. The optimizer, learning rate, and epoch value were then set. Next, data training was conducted. For reproduction, the random seed was fixed, and the gradient was initialized. During this process, a loss function was set for accuracy calculation.

Next, the accuracy was obtained by multiplying the value by 100; then, training was performed repeatedly according to the number of epochs. Data were imported repeatedly based on the batch setting size in the DataLoader. Then, the batch was sent to the GPU, where data were extracted, and the loss value was obtained. Finally, the average and total loss values were calculated.

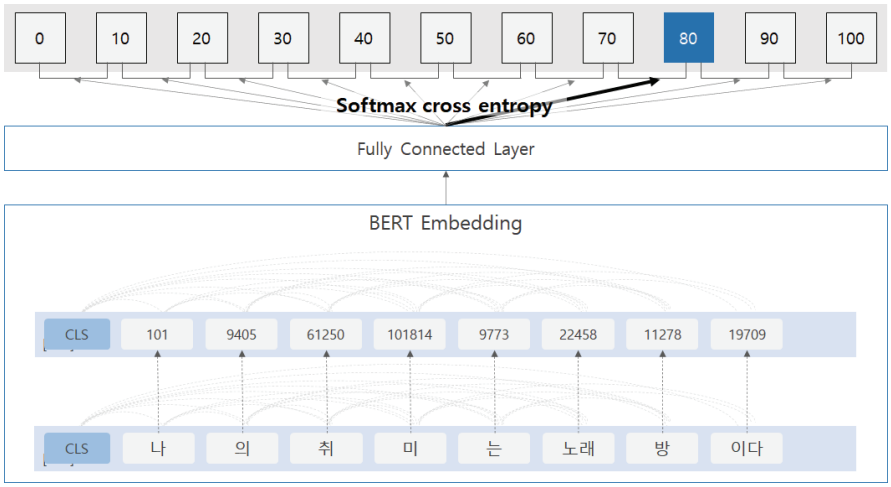

Fig. 3. Fine tuning model for automatic scoring.

4. Performance Analysis

4.1 Experimental Environment

The BERT program used in this paper for automatic scoring was implemented using Windows 10 with 16-GB memory, Python 3.6, TensorFlow 2.2, PyCharm, and a Google Colab GPU Tesla T4. The Adam-W optimizer was used with a learning rate of 0.00003 to set up the generative adversarial network (GAN) experiment. The contents related to the experimental environment are shown in Table 5.

Table 5. Experimental environment

		Value
Subjects and contents	Foreign students	586 participants
	Topic 1 (my hobby)	371 writing samples
	Topic 2 (memorable travel)	215 writing samples
H/W	GPU	Colab Tesla T4
	RAM	16 GB
	OS	Window 10
S/W	Language	Python 3.6
	Framework	TensorFlow 2.2

The answers, including scores and labels, were used for automatic scoring. Among the descriptive answers written by foreign learners, a total of 586 answers were used, which included 371 corresponding to “my hobby” and 215 answers under the topic of “memorable travel.” The data were divided into 70% versus 30% portions: one for training and the other for testin0g, respectively. During the verification stage, 10% of the training data were used.

For the WScore model creation, the BERT model utilizing GPU for classification was created. The optimizer was set to AdamW, and the learning rate was set to 0.00003 with an epoch size of 40. The hyperparameters of the BERT-based WScore are shown in Table 6.

Table 6. Hyperparameters of the BERT-based WScore

Parameter	Value
optimizer	AdamW
max_length	384
batch_size	4
ir	0.00003
epoch	40

4.2 Result of the Experiment

For model learning, accuracy calculation and loss functions were set to reflect the order of scores. Next, a third accuracy was obtained by multiplying the value by 100; then, training was performed repeatedly based on settings. The data loader repeatedly moved data of the respective batch sizes into the GPU. The data were extracted from the batch to calculate the average loss value. Through this process, the accuracy was confirmed as follows. As shown in Table 7, the accuracy was 84.49% at one epoch, and when it reached 20 epochs, the last accuracy was 93.62%. It was the highest at 40 epochs. It was confirmed that, even if the number of epochs was increased to 50 and 100, the accuracy no longer increased.

Table 7. Accuracy by epoch

	epoch 1/40	epoch 20/40	epoch 40/40
Average training loss	0.07	0.00	0.00
Training epoch took	0:00:22	0:00:15	0:02:16
Accuracy (%)	84.49	93.64	95.80
Validation took	0:00:01	0:00:00	0:00:04

5. Conclusion

This paper used a BERT transfer language model to apply an automatic scoring standard to Korean-learning writing tests for foreign learners to assess their efficacy. The data included score labels attributed to descriptive answers written by foreign learners, including their test scores. 371 answers were assessed under the topic, “My hobby,” and 215 were assessed under the topic, “Memorable travel.” For a total of 586 subjects, the training and testing datasets were divided 70 vs. 30%, and 10% of the training data were used for verification. For the automatic scoring of descriptive answers, the BERT-based WScore model was developed, achieving 95.80% accuracy, which is high compared with the methods assessed in previous studies. The most important part of the proposed descriptive scoring method depends on the scoring consistency, which is judged based on valid criteria. Hence, if these scoring criteria and processes can be automated, scoring costs and the temporal burdens of large-scale evaluations can be greatly reduced. This study provides the opportunity to further improve automatic scoring capabilities that analyze foreign learners’ written answers. In the future, the efficiency and consistency of writing test scoring will benefit by establishing large-scale learner datasets and setting standards for Korean language evaluation. However, pertaining to the scoring of the formative evaluation dimension of writing education, student feedback should be centered. Hence, modified approaches are needed.

References

[1] S. H. Ahn and C. S. Kim, "A study on the features of writing rater in TOPIK writing assessment," *Journal of Korean Language Education*, vol. 28, no. 1, pp. 173-196, 2017.

[2] S. Hwang and K. Kim, "BERT-based classification model for Korean documents," *Journal of Society for e-Business Studies*, vol. 25, no. 1, pp. 203-214, 2020.

[3] J. O. Min, J. W. Park, Y. J. Jo, and B. G. Lee, "Korean machine reading comprehension for patent consultation using BERT," *KIPS Transactions on Software and Data Engineering*, vol. 9, no. 4, pp.145-152, 2020.

[4] C. H. Lee, Y. J. Lee, and D. H. Lee, "A study of fine tuning pre-trained Korean BERT for question answering performance development," *Journal of Information Technology Services*, vol. 19, no. 5, pp. 83-91, 2020.

[5] K. Jiang and X. Lu, "Natural language processing and its applications in machine translation: a diachronic review," in *Proceedings of 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)*, Chongqing City, China, 2020, pp. 210-214.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, 2019, pp. 4171-4186.

[7] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016, pp. 715-725.

[8] J. E. Kim, K. Park, J. M. Chae, H. J. Jang, B. W. Kim, and S. Y. Jung, "Automatic scoring system for short descriptive answer written in Korean using lexico-semantic pattern," *Soft Computing*, vol. 22, no. 13, pp. 4241-4249, 2018.

[9] National Institute of the Korean Language, *Application Research of Korean Language Curriculum*. Seoul, Korea: National Institute of the Korean Language, 2017.

[10] J. H. Lee, "A study on error determination standard and classification in Korean education," *Journal of Korean Language Education*, vol. 13, no. 1, pp. 175-197, 2002.

[11] Y. Oh, "BERT with SentencePiece for Korean Text," 2020 [Online]. Available: https://github.com/yeontaek/BERT-Korean-Model.

[12] H. Lee, J. Yoon, B. Hwang, S. Joe, S. Min, and Y. Gwon, "KoreALBERT: pretraining a Lite BERT model for Korean language understanding," in *Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 5551-5557.

[13] Test of Proficiency of Korea (TOPIC) [Online]. Available: https://www.topik.go.kr/HMENU0/HMENU00018.do.

Jung Hee Lee **https://orcid.org/0000-0001-8482-1065**

She received the B.S. degree in Korean Language and Literature and the M.S. degree in Korean Language Education, and Ph.D. degree in Korean Language and Literature from Kyung Hee University, Seoul, Korea. She is Professor in the Graduate School of Education at the Kyung Hee University, where she teaches Korean language Education as a Foreign Language. She specializes in Korean language pedagogy and her research interests include Korean language assessment as a foreign Language, Korean Language Curriculum and Material development. She has been researching as Visiting Scholars at Georgetown University at Washington D.C., USA, in 2012.

Ji Su Park **https://orcid.org/0000-0001-9003-1131**

He received his B.S. and M.S. degrees in Computer Science from Korea National Open University, Korea, in 2003 and 2005, respectively and Ph.D. degree in Computer Science Education from Korea University, 2013. He is currently a Professor in Dept. of Computer Science and Engineering from Jeonju University in Korea. His research interests are in mobile grid computing, mobile cloud computing, cloud computing, distributed system, computer education, and IoT. He is employed as managing & associate editor of Human-centric Computing and Information Sciences (HCIS) by Springer, The Journal of Information Processing Systems (JIPS) & KIPS TRANSACTIONS ON SOFTWARE AND DATA ENGINEERING by KIPS He has received "best paper" awards from the CSA2018 conferences and "outstanding service" awards from CUTE2019 and BIC2020. He has also served as the chair, program committee chair or organizing committee chair at several international conferences including World IT Congress, MUE, FutureTech, CSA, CUTE, BIC.

Jin Gon Shon **https://orcid.org/0000-0002-0540-4640**

He received the B.S. degree in Mathematics and the M.S. and Ph.D. degrees in Computer Science from Korea University, Seoul, Korea. Since 1991, he has been with the Department of Computer Science, Korea National Open University (KNOU). He had been researching as Visiting Scholars at State University of New York (SUNY) at Stony Brook, USA, in 1997, at Melbourne University, Australia, in 2004, and Indiana University, USA, in 2013. After serving the Head of Information & Computer Center and the Head of e-Learning Center, Professor Shon had established the Department of e-Learning in KNOU, offering the first master program of e-Learning in Korea, and served as the Chair of the Department until 2010. He had also worked for KNOU as Director of the Digital Media Center, where all of KNOU e-learning contents and TV programs are produced. Since 1991, he has been working as well for the community services, as chairs or members in various committees including a Vice President of Korea Information Processing Systems and a Vice President of e-Learning Society. His research interests are mainly focused on computer networks, modeling & simulation, distributed computing, wireless sensor networks, e-learning, and especially in ITLET (Information Technology for Learning, Education, and Training) as a member of Korean Delegation to ISO/IEC JTC1/SC36 since 2000. He has made presentations in many conferences, and he won a few of Best Paper Awards including the Gold Medal Paper in the 24th AAOU Annual Conference in 2010. He has also published over 40 scholarly articles in the noted journals and written several books on computer science and e-learning.