논문 2024-2-5 http://dx.doi.org/10.29056/jsav.2024.06.05

Modeling Short Answer Grading Performance Improvement by GPT Augmentation Data

Maresha Caroline Wijanto*, Hwan-Seung Yong*†

Abstract

The automatic grading of short answer question is important in the field of Natural Language Processing. ASAG (Automated Short Answer Grading) task have undergone numerous advancements. Recent studies have adopted transformer models such as the T5 embedding or BERT-base models. Nonetheless, ASAG tasks encounter significant challenges stemming from limited data availability. The urgent need for more training data emerges as a central issue. Several researchers have proposed augmentation approaches to address this gap. In this study, we introduce other data augmentation technique utilizing prompt engineering by the GPT model. We deploy ASAG system using the Sentence Transformers model, fine-tuning specific hyper-parameters alongside the augmented dataset. The primary factors influencing performance enhancement include the augmentation process, particularly the quantity of augmented data, and the dataset split size for training and testing purposes. Furthermore, alternative GPT models or fine-tuning GPT could be explored within the augmentation process.

keywords : Data Augmentation, Automated Short Answer Grading System, GPT, fine-tuning

1. Introduction

The automatic grading of short answers in open-ended questions is a key focus within the field of the Natural Language Processing (NLP) domain. With ongoing developments in NLP and machine learning, educators are increasingly intrigued by the idea of creating exams included open-ended questions that can be automatically graded for large groups of students[1]. Short answer questions are considered semi-open-ended rather than fully open-ended[2]. Typically, two primary assessment methods are preferred for automatically grading open-ended questions. The first approach relies on predefined criteria for assessment, whereas the second method evaluates responses by their semantic similarity to the correct answer. This method compares student responses to a reference correct answer, evaluating them based on how closely they align in terms of semantic meaning[1].

* Computer Science and Engineering, Department of Artificial Intelligence and Software, Ewha Womans University
† Corresponding Author: Hwan-Seun Yong(email: hsyong@ewha.ac.kr)
Submitted: 2024.05.16. Accepted: 2024.06.01. Confirmed: 2024.06.20.

Short answers can be described as texts that adhere to the following guidelines[3, 4]:

- Student's answer to a given question should be expressed in natural language
- Student answer's length should vary from a single phrase to a paragraph
- Student's answer should show knowledge acquired beyond the explicit content of the question
- Grading should be based on objective criteria related to the quality of the content, rather than subjective assessments of writing style.

The ASAG (Automated Short Answer Grading) task has seen numerous advancements, ranging from traditional approaches to classical machine learning and deep learning methods[4, 5]. More recently, deep learning methodologies, including transformer models, have gained traction in ASAG system development. Research by Gomaa, et al. utilize transformer models such as T5-XL embedding model, BERT-base, and all-distilroberta-v1 in the embedding phase[6]. In other research, an SBERT architecture with a pre-trained language model (PLM) is used for training. The SBERT model fine-tunes a pre-trained transformer (BERT) to yield useful sentence embeddings. The experimentation has used stsb-distilbert-base, paraphrase-albertsmall-v2, and quora-distilbert-base pre-trained sentence transformer models[7]. Previous research in the ASAG domain has explored various strategies, including traditional linguistic features coupled with statistical models and neural approaches. Among these possible approaches, the appropriate one in a particular scoring condition is largely dependent on the availability of a manually annotated, question-specific dataset. Studies have shown promising performance when some number of human-scored answers are accessible for each question as training data[8].

ASAG tasks face significant challenges due to limited data availability in many domains[7]. Some researchers confirmed that the performance of the ASAG systems is dependent on the amount of training data[7, 9]. The urge to obtain more training data is the key to the current problem. Augmentation techniques increase training data volume, consequently enhancing model performance. Despite this, the role of data augmentation to improve short answer grading within ASAG system's research remains relatively unexplored. Bonthu, et al. have proposed five augmentation approaches, include Random Deletion, Synonym Replacement, Random Swap, BackTranslation, and using NLPAug Library[7].

Paraphrase generation techniques can be categorized into two primary groups: controlled paraphrase generation methods and deep learning methods. Controlled methods rely on handwritten rules and thesaurus-based alignments or use Statistical Machine Translation (SMT-based) techniques for paraphrase generation. Inspired by the achievements of deep learning networks, paraphrase systems use available parallel corpora to train sequence-to-sequence models, aiming for enhanced performance. Recently,

there has been a surge in the use of large language models employing transformer architectures and less supervised data across various NLP tasks. Consequently, some researchers have adopted these model frameworks and adapted their code for diverse NLP tasks, including paraphrasing[10].

In this research, we introduce a data augmentation method aimed at enhancing the performance of an ASAG system. Our approach involves employing GPT as a strategy to supplement our dataset with additional data.

GPT-4 and GPT-3.5 exhibit numerous similarities, primarily relying on a comparable Transformers architectural model, albeit at different scales. GPT-4 surpasses GPT-3.5 significantly in size, boasting 170 trillion parameters compared to GPT-3.5's 175 billion parameters. This substantial size discrepancy underscores enhanced capabilities in performance and accuracy, particularly in managing complex language models and natural language processing tasks[11].

In addition to considering dataset size, our approach involves utilizing several of the latest Pre-trained Language Models, specifically BERT-based models, during the training phase of the recommended ASAG system. Furthermore, we will fine-tune certain hyper-parameters to optimize our results.

We divide this paper into the following sections. Section 2 reviews all the work related to Data Augmentation in ASAG. Section 3 presents the proposed methods, including the datasets, evaluation metrics, and experiment settings used in this research. Section 4 presents the implementation of the system and discusses the results. Section 5 summarizes all the achievements from these experiments.

2. Data Augmentation in ASAG

Lun, et al. introduced MDA-ASAS[12], a method comprising multiple data augmentation strategies aimed at enhancing performance in automatic short answer scoring. MDA-ASAS is designed to refine language representation through diverse augmentation methods, such as back-translation, utilizing correct answers as reference points, and content swapping. They argue that external knowledge significantly influences the ASAS process. Simultaneously, the effectiveness of the Bidirectional Encoder Representations from Transformers (BERT) model in improving various natural language processing tasks has been well-documented. BERT harnesses extensive unsupervised data to acquire semantic, grammatical, and other relevant features, effectively integrating external knowledge. By leveraging the latest BERT model, their experimental findings on the ASAS dataset demonstrate that MDA-ASAS yields substantial improvements over existing methodologies. Specifically, in the 5-way comparison, both accuracy and weighted-average-F1 metrics outperform all other methods.

The second approach proposed to enhance ASAG performance involves transfer learning and augmentation[7]. It entails fine-tuning

three sentence transformer models on the SPRAG (Short Programming Related Answer Grading Dataset) corpus, and employing five augmentation techniques: Random Deletion, Synonym Replacement, Random Swap, Backtranslation, and using NLPAug Library. The SPRAG corpus featuring keywords and special symbols, totaling 4039 records and constituting a binary classification task. Experimentation involves varying data sizes (25%, 50%, 75%, and 100%) with augmented data to assess the influence of training data on the fine-tuned sentence transformer model. Training utilizes an SBERT architecture with a pre-trained language model (PLM). The study utilizes stsb-distilbert-base, paraphrase-albertsmall-v2, and quora-distilbert-base pre-trained sentence transformer models. This research offers a comprehensive examination of fine-tuning pre-trained sentence transformer models using different data sizes through text augmentation techniques. Results indicate that employing random swap and synonym replacement concurrently during fine-tuning significantly enhances performance, with a 4.91% accuracy increase (reaching 84.21%) and a 3.36% increase in F1-score (reaching 88.11%).

A recent study[1)], published in 2024, introduces paraphrase generation and supervised learning techniques to enhance ASAG performance[10]. Initially, they present a sequence-to-sequence deep learning model aimed at generating plausible paraphrased reference answers based on the provided reference answer. Additionally, they proposed a supervised grading model based on sentence embedding features, which enriches features to enhance accuracy by considering multiple reference answers. Experiments are conducted both in Arabic and English. They show that the paraphrase generator produces accurate paraphrases. Using multiple reference answers, the proposed grading model achieves a Root Mean Square Error (RMSE) of 0.6955 and a Pearson correlation of 88.92% for the Arabic dataset, and an RMSE of 0.779 and a Pearson correlation of 73.5% for the English dataset.

1) https://github.com/DigiKlausur/ASAG-Dataset

3. Proposed Method

In this section, we will describe our proposed method regarding the augmentation of dataset and the short answer grading system.

3.1 Dataset

This research utilized Assisted Automated Short Answer Grading Dataset[1], which comprises data from an examination in a neural network course. The course was taken by graduate students at the University of Applied Sciences Bonn-Rhein-Sieg. Student's answers were collected through Jupyter notebooks. A total of 38 students participated in the examination, with each exam consist of 17 questions. Thus, the dataset contains a total of 646 question answers. These responses were evaluated by a single human judge, who

assigned scores on 3-way, describe as an integer scale ranging from 0 (completely incorrect), 1 (partially correct), and 2 (perfect answer). Table 1 shows the row amount of data for each grade label.

Table 1. ASAG dataset grade label distribution

Grade Label	Number of Data
2	333
1	219
0	54

Table 2 shows an example of dataset that illustrates the question, the desired answer as a correct answer, and the student answers with the existing grade label range.

3.2 Proposed Method

To address the issue of data imbalance, we propose a data augmentation approach utilizing GPT. Specifically, we utilize two GPT models in this study: GPT-3.5 (model: gpt-3.5-turbo-1106) and GPT-4 (model gpt-4). The approach involves prompt engineering using GPT to generate new sentences for each grade label. We implement the prompt engineering based on specific characteristics corresponding to each grade label:

- For label 0: Generate new sentences opposite to the desired answer in the dataset
- For label 1: Paraphrase existing student answers to generate new sentences. The quantity of data depends on the existing number of data and the maximum amount of data in other labels
- For label 2: Paraphrase existing desired answers to generate new sentences.

Table 2. Example dataset

Question: Explain the Bias Variance Dilemma!	
Desired Answer: Bias-variance dilemma is a principle supervised learning problem. The dilemma arises due to the variance of data and bias of model. When there is high bias, the model fits the training data perfectly but suffers from high variance, when the bias is low the variance reduces but the model doesn't fit the data well. This dilemma makes the generalizability difficult to achieve.	
Student Answer	Label
Usually only one of Bias and Variance can be minimized. In an RBFN for example few kernels with greater width leads to a high bias but a low variance. If you choose many kernels with smaller width the bias is low but the variance is high. Higher complexity models need more training data.	2
Ideally bias and variance would be 0 after learning a machine. However, bias and variance counteract each other: when bias decreases, variance rises and respectively in the other direction. This leads to the dilemma that either one of the values has to be present.	1
Bias is provides an affine transformation. and it is treated as extra inputs. which normally taken as +1	0

The prompt text implemented in the system depends on the grade label. For grade label 0, the prompt text will be "Please make a completely different sentence from this following sentence: '{answer}' so it counts as an opposite sentence" to get the opposite sentences. While for the grade label 1 and 2, the prompt text will be "Please paraphrase the following sentence '{answer}'" to get similar sentences.

By constructing appropriate prompts tailored to the paraphrasing task, we leverage the advanced natural language processing capabilities of GPT-3.5 and GPT-4 to generate a diverse range of linguistically and contextually relevant rephrasing of student answers.

In this research, we also proposed a pre-trained language model SBERT with hyper-parameters fine-tuning to automatically grading the short answer. As other research found out that the best result that can compete other research came from the all-distilroberta-v1[13], we will also perform the grading system using that model. We will apply the new augmented dataset with the all-distilroberta-v1 model and the best hyper-parameter combination.

The hyper-parameter combination include some fixed and also adjustable hyper-parameters. The fixed hyper-parameter values including the pre-processing step to remove the special characters and change them into lowercase, using gradient checkpointing to reduce memory usage, setting the number of epochs to 12, and the batch size to 16. Then, for the adjustable hyper-parameter, we will also check whether removing stop words and differences in size between the training and testing data splits will affect the performance results.

3.3 Evaluation Metrics

In this study, we used some evaluation metrics, including RMSE, Accuracy, Pearson Correlation and Cosine Similarity for data augmentation. RMSE, or Root Mean Square Error, is a common measures used to evaluate the quality of prediction using Euclidean distance. Accuracy measures the percentage of correctly graded answer, but it has limitations in scenarios with imbalanced dataset[14]. While Pearson Correlation used to evaluate the strength and presence of a linear relationship between the predicted and manual grades[6].

4. Result and Discussion

4.1 Data Augmentation

Based on the scenario mentioned above, we augmented the dataset using GPT with a temperature value of 0.7. The temperature value refers to the degree of randomness of the newly generated text. Fig. 1. shows the results of the new dataset after augmentation process. We attempted to double the size of the dataset to assess whether the amount of data also affects the performance of the ASAG system.

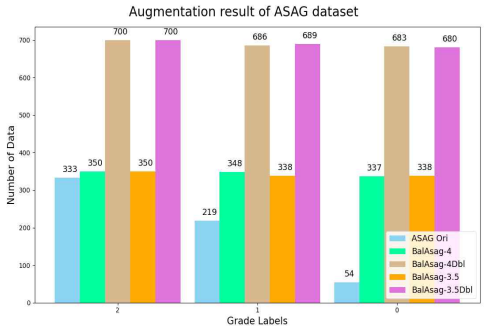

Fig. 1. Number of data after augmentation

The total amount of data for each augmentation process using GPT-4.0 (BalASAG-4), double size GPT-4.0 (BalASAG-4Dbl), GPT-3.5 (BalASAG-3.5), and double size GPT-3.5 (BalASAG-3.5Dbl) is 1035 rows, 2069 rows, 1026 rows, and 2062 rows, respectively. We checked the similarity score between the desired answer and the students' answer for both the original and the newly generated text. The cosine similarity score can be seen in Table 3.

Table 3. ASAG dataset grade label distribution

Dataset	Label 0	Label 1	Label 2
BalASAG-3.5	0.4884	0.6558	0.7399
BalASAG-3.5Dbl	0.4869	0.692	0.792
BalASAG-4	0.4567	0.684	0.7379
BalASAG-4Dbl	0.4565	0.6984	0.8051

We observe that with more data, the similarity score also increases. Additionally, data augmentation using GPT-4.0 (BalASAG-4) generally yields better results. For the 0-grade label, a smaller value indicates better performance as it should contain answers that are furthest away from the desired answer or correct answer.

4.2 ASAG Result

After obtaining the newly augmented dataset, we implement it into our grading system model. We utilize the all-distilroberta-v1 model along with some fixed parameters. Then, we analyze the performance using the evaluation metrics mentioned above. Fig. 2. displays the results

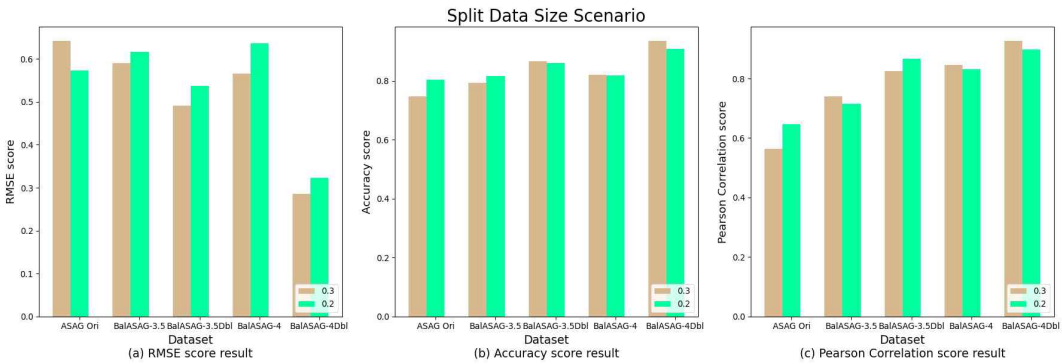

Fig. 2. Performance result by dataset split size scenario

based on RMSE, accuracy and Pearson correlation scores for different training-testing data split scenarios. We employ data splits with sizes 0.2 and 0.3, indicating that 20% or 30% of the data will be used for testing, while the remaining will be used for training.

A smaller RMSE score indicates better performance, whereas higher accuracy or Pearson correlation scores reflect better the performance. In general, the best results were obtained from BalASAG-4Dbl dataset and a data split size of 0.3. The double-sized dataset also demonstrate a steady increase across all scenarios. Moreover, when compared with the original data, other scenarios also show improvements for all existing evaluation metrics.

Based on the results in Fig. 2a., we observe an improvement in the RMSE score, from 0.6427 for the original dataset to 0.2861 for the BalASAG-4Dbl dataset. The same trend is also seen for the accuracy score in Fig. 2b, with the original dataset scoring 0.7473, which increases to 0.9357 for the BalASAG-4Dbl dataset. Similarly, for dataset undergoing the same augmentation process, the Pearson Correlation score increases from 0.5644 for the original dataset to 0.9273 This can be seen in Fig. 2c.. These findings are based on a data split size scenario for training-testing of 0.3, indicating a significant increase in performance from the augmentation process.

We do some additional experiments, checking whether removing stop words from the dataset will make performance result better or not. We implement this scenario using all-distilroberta-v1 model, gradient checkpointing, and dataset split size=0.3. We will first discuss the performance results of the model that also applies removing stop words. The results are shown from the second bar chart in Fig. 3. All RMSE scores for all datasets from the augmentation process experienced an improvement, falling below 0.6745 for the original dataset. The smallest RMSE value of 0.4478 was obtained from the BalASAG-4Dbl dataset, that can be seen in Fig. 3a. The same trend also happened for accuracy and Pearson correlation score, where the original dataset only scored 0.7253 for

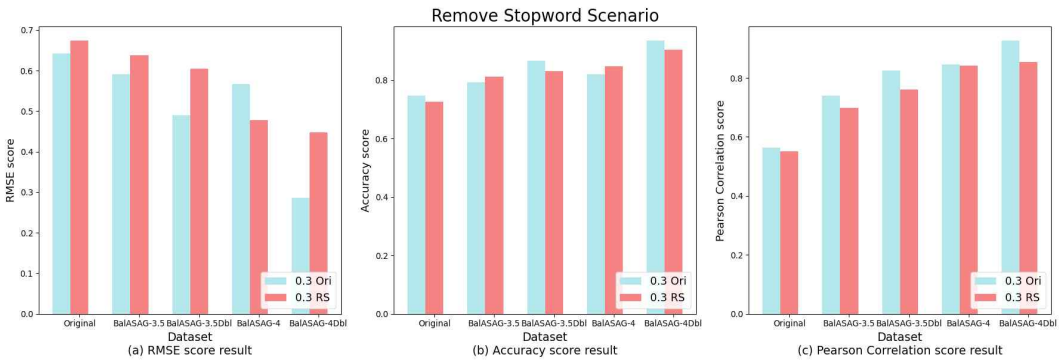

Fig. 3. Performance result by remove stopwords scenario

accuracy and 0.5507 for Pearson Correlation. The detail result of accuracy shown in Fig. 3b, while the result of Pearson Correlation score shown in Fig. 3c. The implementation of the augmented dataset all showed better results than the original dataset. Particularly, the performance results from the double-sized dataset resulting from the augmentation process using GPT-4.0 reached 0.9035 for accuracy score and 0.8538 for Pearson Correlation. From these two experiments we can also see that the use of a double-sized augmented dataset improves the ASAG model performance.

However, also based on Fig. 3., we can see that the performance of the original dataset is better than the dataset after removing stop words. This may be caused by fewer words being processed. This system is a short answer grading system, so the text provided is not too long. When some words are removed from the available sentences, there will be less data that can be processed. When data is reduced, there is also the possibility of changes in meaning.

For RMSE score, the best score from the original dataset reached 0.2861, while for the results of implementing removing stop words, the RMSE score only reached 0.4478. The accuracy score also experiences the same phenomenon. The best accuracy score from the original dataset reached 0.9357, while with removing stop words implementation, it only reached 0.9035. The Pearson Correlation score is also the same. The best result when implementing removing stop words is only 0.8538, whereas the original dataset can reach a correlation score of 0.9273.

5. Conclusion

In this paper, we proposed an augmentation process for dataset to enhance the performance of short answer grading system. This augmentation process utilizes prompt engineering from GPT, specifically GPT-4.0, which produces better similarity score values. Based on experimental results, double-sized datasets provide the best performance results. This short answer grading system applies a pre-trained Sentence Transformers model, particularly all-distilroberta-v1 and by applying appropriate fine-tuning hyper-parameter, the system achieves the best performance with RMSE, accuracy and Pearson Correlation scores reaching 0.2861, 0.9357, and 0.9273, respectively. However, the additional process of removing stop words did not show an improvement in system performance. The main factors affecting performance improvement are augmentation process, specifically the amount of augmented data, as well as the dataset split size for training and testing data. Furthermore, alternative GPT models or fine-tuning the GPT API could be explored in the augmentation process.

References

[1] C. N. Tulu, O. Ozkaya, and U. Orhan,

"Automatic Short Answer Grading with SemSpace Sense Vectors and MaLSTM," IEEE Access, vol. 9, pp. 19270 - 19280, 2021, doi: 10.1109/ACCESS.2021.3054346.

[2] Z. Zhang, E. Strubell, and E. Hovy, "A Survey of Active Learning for Natural Language Processing," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.10109

[3] A. Ahmed, A. Joorabchi, and M. J. Hayes, "On Deep Learning Approaches to Automated Assessment: Strategies for Short Answer Grading," in International Conference on Computer Supported Education, CSEDU - Proceedings, Science and Technology Publications, Lda, 2022, pp. 85 - 94. doi: 10.5220/0011082100003182.

[4] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," International Journal of Artificial Intelligence in Education, vol. 25, no. 1. Springer New York LLC, pp. 60 - 117, Jan. 10, 2015. doi: 10.1007/s40593-014-0026-8.

[5] S. Haller, A. Aldea, C. Seifert, and N. Strisciuglio, "Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2204.03503

[6] W. H. Gomaa, A. E. Nagib, M. M. Saeed, A. Algarni, and E. Nabil, "Empowering Short Answer Grading: Integrating Transformer-Based Embeddings and BI-LSTM Network," Big Data and Cognitive Computing, vol. 7, no. 3, Sep. 2023, doi: 10.3390/bdcc7030122.

[7] S. Bonthu, S. Rama Sree, and M. H. M. Krishna Prasad, "Improving the performance of automatic short answer grading using transfer learning and augmentation," Eng Appl Artif Intell, vol. 123, Aug. 2023, doi: 10.1016/j.engappai.2023.106292.

[8] S.-Y. Yoon, "Short Answer Grading Using One-shot Prompting and Text Similarity Scoring Model," May 2023, [Online]. Available: http://arxiv.org/abs/2305.18638

[9] D. Wilianto and A. S. Girsang, "Automatic Short Answer Grading on High School's E-Learning Using Semantic Similarity Methods," TEM Journal, vol. 12, no. 1, pp. 297 - 302, Feb. 2023, doi: 10.18421/TEM121-37.

[10] L. Ouahrani and D. Bennouar, "Paraphrase Generation and Supervised Learning for Improved Automatic Short Answer Grading," Int J Artif Intell Educ, 2024, doi: 10.1007/s40593-023-00391-w.

[11] A. Koubaa, "GPT-4 vs. GPT-3.5: A Concise Showdown," TechRxiv, Apr. 2023, doi: https://doi.org/10.36227/techrxiv.22312330.v2

[12] J. Lun, J. Zhu, Y. Tang, and M. Yang, "Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring," in The AAAI Conference on Artificial Intelligence, 2020, pp. 13389 - 13396. doi: https://doi.org/10.1609/aaai.v34i09.7062.

[13] M. C. Wijanto and H.-S. Yong, "Combining Balancing Dataset and SentenceTransformers to Improve Short Answer Grading Performance," submitted to Applied Science, 2024.

[14] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. The MIT Press, 2016. Accessed: Apr. 16, 2024. [Online]. Available: http://www.deeplearningbook.org

Authors

Maresha Caroline Wijanto

2011.8-2013.10 Master of Informatics from Institut Teknologi Bandung, Indonesia
2021.9-present Ph.D student in Computer Science and Engineering, Ewha Womans University, South Korea
2010.2-present Faculty member in Maranatha Christian University, Indonesia
<Research interests> Data Mining, Machine Learning, Natural Language Processing

Hwan-Seung Yong

1994.2 Ph.D in Computer Engineering from Seoul National University
1985.2-1989.2 ETRI Research Member
2002.9-2003.2 IBM T.J.Watson Research Lab. Visiting Scholar
1995.3-present Professor in Ewha Womans University
<Research interests> Databases, Data Mining, Articial Intelligence