Journal of Digital Convergence
Vol. 19. No. 9, pp. 181-188, 2021

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2021.19.9.181

Tax Judgment Analysis and Prediction using NLP and BiLSTM

Yeong-Keun Lee[1], Koo-Rack Park[2*], Hoo-Young Lee[3]
[1]Ph.D, Dept. of Computer Engineering, Kongju National University,
[2]Professor, Dept. of Computer Science & Engineering, Kongju National University,
[3]Ph.D, Dept. of Computer Engineering, Kongju National University

NLP와 BiLSTM을 적용한 조세 결정문의 분석과 예측

이영근[1], 박구락[2*], 이후영[3]
[1]공주대학교 컴퓨터공학과 박사, [2]공주대학교 컴퓨터공학부 교수, [3]공주대학교 컴퓨터 공학과 박사

Abstract Research and importance of legal services applied with AI so that it can be easily understood and predictable in difficult legal fields is increasing. In this study, based on the decision of the Tax Tribunal in the field of tax law, a model was built through self-learning through information collection and data processing, and the prediction results were answered to the user's query and the accuracy was verified. The proposed model collects information on tax decisions and extracts useful data through web crawling, and generates word vectors by applying Word2Vec's Fast Text algorithm to the optimized output through NLP. 11,103 cases of information were collected and classified from 2017 to 2019, and verified with 70% accuracy. It can be useful in various legal systems and prior research to be more efficient application.

Key Words : Artificial Intelligence, Legal System, Tax Tribunal, Word2Vec, BiLSTM, NLP.

요 약 일반인에게 난해한 법률분야를 이해하기 쉽고 예측 가능 할 수 있도록 인공지능을 적용한 법률 서비스에 대한 연구의 중요성이 대두되고 있다. 본 연구에서는 조세심판원의 결정정보를 수집하고 데이터 처리와 자체 학습을 통한 모델을 구축하여 사용자의 질의에 맞는 답변을 예측하기 위한 시스템을 제안한다. 제안 모델은 웹크롤링을 통해서 조세 결정문의 정보 수집 및 자연어 처리과정을 통하여 유용한 데이터를 추출하고, 최적화된 산출물을 Word2Vec의 Fast Text 알고리즘을 적용하여 단어의 벡터를 생성하였다. 2017년부터 2019년까지 총 11,103건의 정보를 수집하고 분류하였으며 RNN 기술의 BiLSTM을 적용하여 자체학습을 통한 결과 예측 프로그램을 구축하여 70%정확도로 실증하였다. 향후 다양한 법률시스템으로 활용성을 기대할 수 있으며 보다 효율적인 적용을 위한 연구와 정확도 향상을 위한 연구가 계속되어야 한다.

주제어 : 인공지능, 법률 서비스, 조세심판원, 자연어처리, Word2Vec, BiLSTM.

*본 논문은 제1저자의 박사학위 논문을 축약한 논문임.

*Corresponding Author : Koo-Rack Park(ecgrpark@kongju.ac.kr)

Received August 31, 2021
Revised September 13, 2021
Accepted September 20, 2021
Published September 28, 2021

1. Introduction

Internet has been wide spread due to the rapid development of the information communication technology, utilization of the artificial intelligence has been discussed in various ways in the legal area thanks to the advancement of the artificial intelligence, the interest on the application of artificial intelligence to the law has been grown from the early stage of the research on the artificial intelligence[1,2], and a lot of efforts have been put into it by scholars of laws and forums such as a legal system these days[3-5]. In addition, in the field of laws related to the artificial intelligence, a lot of efforts have been put into applying it into the legal area, and studies to improve gradually are underway [6]. Legal inference ultimately transfers the practice of the law, and it becomes the foundation for all actions of studies[7]. The core of the law is the cognitive aspect of a human regarding the legal inference[8]. Therefore, the artificial intelligence and the law are subfields of the artificial intelligence study that focus on the design of the computing model and the program of a computer, and it is to simulate or perform the legal inference[9]. The services such as the legal information search, Electronic Discovery (E-Discovery) and legal consultation utilizing the artificial intelligence are being provided these days which have been provided by lawyers, but in reality, it is a difficult to execute diverse legal services using the artificial intelligence. Ultimately, the area of AI-based legal bots is one of technology that would be able to be continuously improved. The development of the artificial intelligence is becoming more important as the solution to resolve various questions of non-experts regarding laws, and therefore the participation of legal experts is mandatory in applying the legal expert system[10]. Also, it is very difficult for general public or non-experts who are not very well aware of laws to interpret laws, and studies on the expert system related to the laws have been insignificant in Korea as it has been difficult to find a joint area with the expert legal knowledge of each area. In this paper, a Taxation Adjudgment Result Prediction Model that learns the points of written taxation adjudgements that have been ruled in recent years through a natural language analysis using the artificial intelligence and infers whether it would be abandoned, admitted or dismissed based on a part of the written judgements that get newly input is proposed. Through the proposed model, it is anticipated that general public and non-experts who have difficulty in understanding the tax laws would be able to approach the law as they would become able to predict the adjudgment result.

2. Related Works

2.1 Word2Vec

Word2Vec uses a word embedding that provides the result regarding a task of processing various languages using the word embedding[11,12], which is widely used in the study of classifying images and text and processing of the natural language. Also, it is based on the distribution hypotheses which states that words with same context are located nearby in a certain embedding space[13]. Word2Vec is a method that can infer the relationship between words since it can simplify the model and reduce the learning time using the technique of reducing the number of hidden layers, and it expresses a word in the form of a vector that has the sequence and the meaning of the word. Therefore, a process of computing the similarity or reducing the dimension is unnecessary, and the word can be quantified through the vector value that includes a distinctive feature. Text data and the data type have a certain language structure that gets limited by the utilization rule and the grammar[14], and accurate embedding for popular items has a very important role in making a recommendation and performing a prediction task[15]. The method that is generally used in Word2Vec that can design, adjust, and create the word embedding in order to process the natural language is SGNS (Skip-gram with Negative Sampling), which is frequently used in creating item embedding from the recommendation. Following Fig. 1 shows two learning algorithms of Word2Vec modeling, and (a) represents CBOW (Continuous Bag-of-Word) and (b) represents Skip-gram[12].

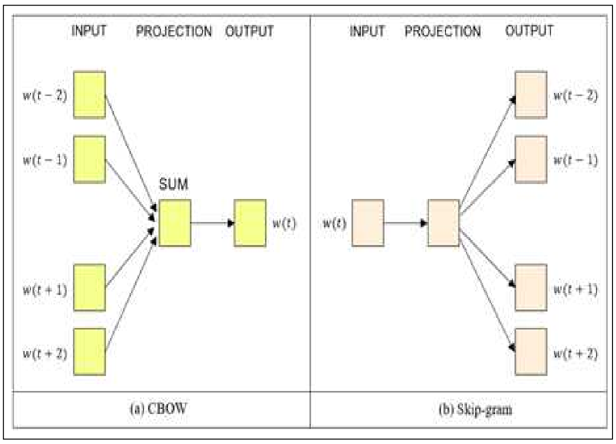

Fig. 1. Two training algorithms of Word2Vec Modeling

However, Word2Vec model can't process a word that doesn't exist in the corpus that has been learned since it has a limitation that the morphologic feature of the word is not reflected very well. Therefore, as the method to overcome such limitation, FastText which was developed by Facebook learns a vector of n-gram rather than a word, and the vector of a certain word is expressed as the average of all n-gram vectors that compose the word[16,17].

2.2 BiLSTM

RNN (Recurrent Neural Networks) is a strong model for variable length sequence data, and it is used to produce a useful result in many tasks including voice recognition and a language model [18] since it is expandable to various learning problems related to the sequential data and is an effective model. Such RNN maintains the memory based on the history information, and it can predict the current output. LSTM (Long Short-Term Memory) is widely used as the architecture of RNN that can process the dependency for a long period of time[19]. LSTM has a weakness that it only uses past context and doesn't use future context. BiLSTM (Bidirectional LSTM) is a method to connect two final output vectors after processing sequences independently in both directions using two independent LSTMs. It is very efficient in the prediction as it can use the context in both directions and is also effective in improving the performance compared to LSTM[20].

3. Proposed System

3.1 Proposed Model Structure

The proposed model is a model that creates a vector of a word using FastText algorithm among Word2Vec which is a technique that collects decision information of a written judgment and displays the word in the vector space, and learns the judgment information regarding the related written judgment using BiLSTM technique. Following Fig. 2 is a system diagram of the proposed model.

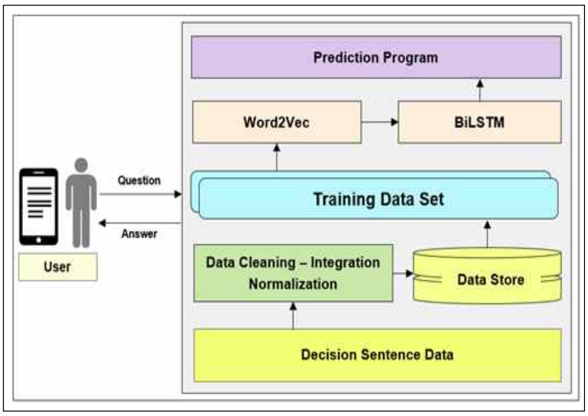

Fig. 2. Proposed Model Structure

The proposed model is composed of a phase that creates data for a practice through the data collection and data normalization; a phase that creates a vector through the data for learning which has been created as the final; a completion of the model with the word vector as the input; and a method to predict the result of a judgment through a sentence that has been entered.

3.2 Collection of Written Judgment Data

The proposed model collects decision point data of a written judgment and performs the process of cleaning the data through the regular expression. As for the written judgment data that was used for this paper, total 11,103 cases that have been decided in three years from 2017 to 2019 in Korea Tax Tribunal to reflect the latest trend of the written judgments were used as the data, and tax items such as corporate tax, income tax, VAT, and transfer tax were used as the data. The relevant cases are determined to be dismissed, abandoned or admitted (cancelled, decided, reinvestigated) according to the result of the hearing. As for each document, the data that is disclosed in the website of Tax Tribunal as its personal information is blocked is collected. Following Fig. 3 is the structure of a recent written judgement of Korea's Tax Tribunal.

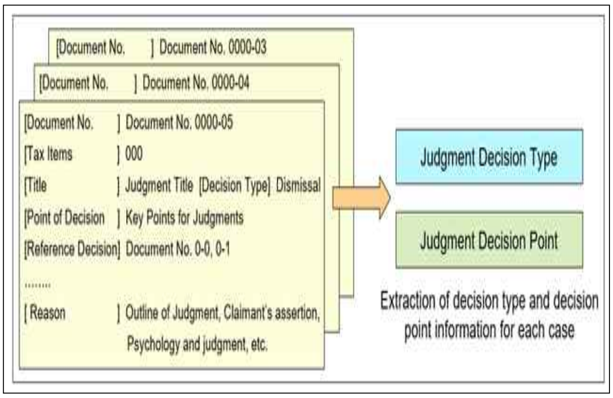
[Document No.] Document No. 0000-03
[Document No.] Document No. 0000-04
[Document No.] Document No. 0000-05
[Tax Items] 000
[Title] Judgment Title [Decision Type] Dismissal
[Point of Decision] Key Points for Judgments
[Reference Decision] Document No. 0-0, 0-1
....
[Reason] Outline of Judgment, Claimant's assertion, Psychology and judgment, etc.
Judgment Decision Type
Judgment Decision Point
Extraction of decision type and decision point information for each case

Fig. 3. Judgment Structure

Written judgment decision includes the standardized items such as document number, tax items, title, point of decision and reference decision, and the proposed model utilized the judgment point data of the written judgment that the judgment decision type and major content of the related judgment are summarized.

Following Table 1 is a partial data regarding decision points of written judgments that have been collected and the decision type information.

Table 1. Data for Learning Written Judgment

No.	Written Decision	Decision Type
1	Since the disposition that is subject to the current request for the judgment doesn't exist as the disposition authority has decided to cancel the imposition disposition of this case by authority, this request for the judgment is determined to be unlawful request.	0
2	It shall be conceded that a house under the size of the national housing that is entitled to the exemption of VAT according to the Special Tax Treatment Control Law means the house that corresponds to the 「Housing Law」 and has the exclusive residential area of 85㎡ or less (100㎡ or less for rural districts such as Eup or Myeon).	2
3	Most of creditors in dispute were investigated as a 'bomb company' that has only issued sales tax invoices without purchasing. The payment that has been paid by a billing company was immediately withdrawn with cash as soon as the payment was deposited to the creditor in dispute, the tax invoice was received before measuring the weight of waste copper, etc.	1

As for the decision type, 'dismissal' is coded to be 0, 'renunciation' is coded to be 1, and 'cancellation', 'reassessment' and 'reinvestigation' which are types of admission are coded to be 2 for learning the result of the written judgment. Each written judgment that has been collected consists of 250 words or less, and legal information that is used for the judgment and the ground for the decision are included. Also, in order to protect personal information that is included in the written judgment from the content that was extracted from the written judgment, the data cleaning that removes identified symbols, punctuation marks, and special characters was performed.

3.3 Creation of Word Vector and Learning Model

Once the collection and cleaning of learning data for a judgment prediction is completed, a task of making the written decision into a word vector is performed. To do the word vectorization task, a task of separating a sentence to units of a word segment is performed using a morpheme analyzer, and during the process, a task of removing a stopword is performed to improve the efficiency of learning. To make a word vector, Genism Library that is generally used was used. As for the algorithm, Fasttext which is a type that utilizes the vector of a partial word of the core word of Word2Vec was used. Fasttext is a model that has higher efficiency regarding the noise data such as the typo or misspelling of the text data compared to Word2Vec. As for the analysis method, skip−gram method that predicts the neighboring word through the core word was used, and appearance frequency of 5 words on the left and 5 words on the right around the neighboring word were computed. Also, the word vector was created as the 300 dimension vector.

The data for the final learning is the text type data that the written judgment and judgment result are stored. As for each written judgment, meaningful word segments got separated through the morpheme analysis, and the words got stored through vectorization using Word2Vec technique. Word vector data classified the written judgment into morphemes and converted it into vector information each, and then the final learning model was created utilizing the BiLSTM which is one of RNN techniques as an input data.

Following Fig. 4 is an overview regarding BiLSTM learning model, and the input data is the 3−dimension type vector with the size of 255 x 300. Rather than learning entire data at once, total 100 times of repetitive learnings were performed by organizing 32 data in the batch structure by dividing the columns of the data, and once the final bi−directional learning is completed, the result gets output as final in the form of Dismiss, Abandon and Admit through SoftMax function by merging the result.

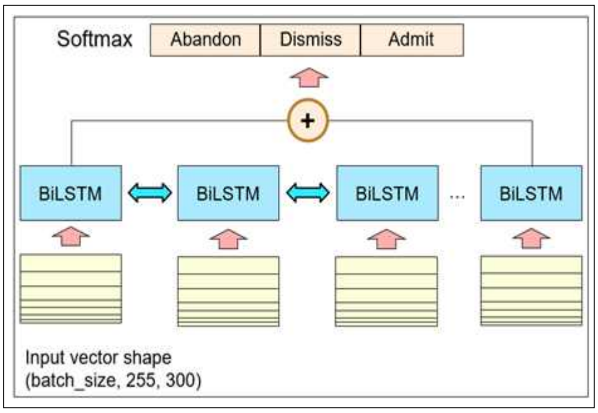

Fig. 4. BiLSTM Learning Model

Following Fig. 5 is the final learning result graph of the prediction model. The prediction model performs the learning in the batch structure, and the learning through the batch was repeated 100 times. As the learning was proceeded, the loss was reduced, but after 40 times, there wasn't significant change in the loss. The final loss was acquired to be 0.003892, which confirms that the learning model is valid.

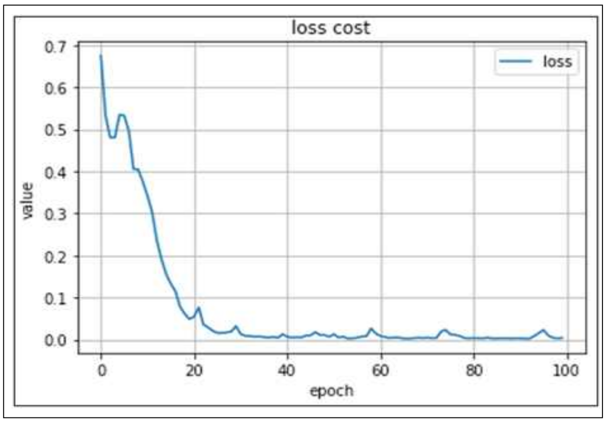

Fig. 5. Final Learning Result of the Prediction Model

4. Results and Discussion

Following Fig. 6 shows the structure of the experiment method that measures the accuracy of the proposed model. As for the prediction program, the experiment to predict the result was performed by inputting major phrase of 10 written judgments, dividing the data that had been input in the units of word segments through the analysis of the morpheme, and inputting these words into the existing prediction model.

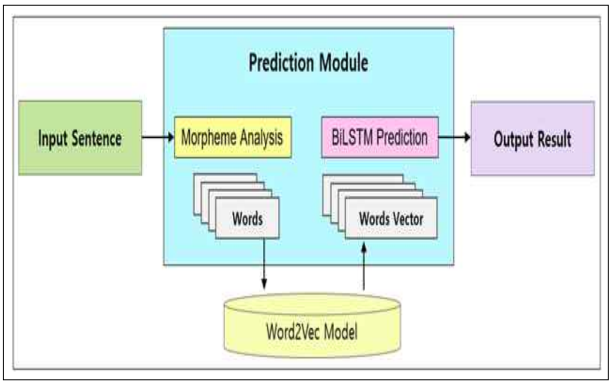

Fig. 6. Structure of Prediction Program for Measuring the Accuracy of the Propose Model

The written judgments shown in Table 2 were input into the model in Fig. 5, and the results were predicted. 10 judgment sentences were input for the experiment, and the validity of the proposed model was confirmed as the incorrect results were predicted from 3 sentences, and the correct results were predicted from remaining 7 sentences as the result of the prediction.

Table 2. Prediction Results through Written Judgments that Have Been Input

No.	Written Judgment Input Sentence	Result (Prediction)	Prediction
1	It can't be stated that the interest can no longer be received as long as it isn't proven that joint sureties to pay investment principal are insolvent.	1(1)	True
2	There is an excessive difference between the value that has been calculated by applying the average transaction price of real transaction prices to the area in dispute and the price of the area in dispute that is evaluated using the officially assessed individual land price that has been applied when the applicant reported the transfer income tax.	1(1)	True
3	Inheritance Tax and Gift Tax have been already paid for the property that was given in advance, and it has been verified that the applicant had returned the amount in dispute to A with the banking evidence after the death of the predecessor	1(2)	False
4	A fact that the applicant received labor income during the time period of cultivating the farm in dispute has been verified, and the applicant has failed to submit the evidences that the applicant had cultivated the farm such as purchasing agricultural pesticides, fertilizer and agricultural materials and the records regarding the harvest of agricultural product.	2(2)	True
5	Considering the entry records of the vehicle that is owned by the corporate, it is difficult to concede that the applicant has resided in house A as the actual space of living with the intention of eingesessen.	1(0)	False
6	The request for the judgment must have been made within 90 days from the data that the applicant became aware of the disposal.	0(0)	True
7	It is difficult to concede that the business right existed since the selling corporates hadn't received the authorization for the project implementation for the real-estate in dispute when the applicant corporate had acquired the real-estate in dispute.	1(1)	True
8	Whereas the applicant bought the land in dispute from A before the division of the land, the contract in dispute states the seller to be B, and therefore the seller in the registration and the seller in the contract in dispute are different.	1(1)	True
9	An objective evidence that can prove that the applicant used the name trust on the stock in dispute can't be verified.	1(2)	False
10	It is a land (baseball field) for sports facility that is defined by Clause 106, Article 1, Paragraph 2 of 「The Local Tax Law」 and Clause 101, Article 3, Paragraph 9 of the enforcement ordinance of the same law, so it can be conceded that the land is actually used as the land for the sport facility.	2(2)	True

5. Conclusion

A lot of changes are being made throughout entire industries due to the development of 4th industrial revolution, the internet has already been widely spread, and the artificial intelligence has improved rapidly throughout entire areas of our society in recent years. Thanks to the improvement of the artificial intelligence, utilization of the related technologies has been discussed in the legal area in various ways. Legal inference ultimately transfers the practice of the law, and it becomes the foundation for all actions of studies. The core of the law is the cognitive aspect of a human regarding the legal inference. Considering such trend, the study to apply the artificial intelligence to the legal area is something that can no longer be delayed, and it is a point of time that the participation of researchers working in the artificial intelligence field and the discussion are necessary. The proposed model which applies NLP (Natural Language) and the artificial intelligence algorithm BiLSTM to written judgment data of last 3 years related to the recent taxation adjudments in Korea predicted the results by referring to the previous precedents, and showed 70% accuracy. But cooperation with legal experts in the processes of raising the accuracy of the prediction and collecting and classifying the related learning data is mandatory. As for the direction of future study, a preceding study regarding the area that looks for various utilization areas for the AI-based legal system and efficiently applies it should be continued.

REFERENCES

[1] J. R. Park & S. O. Noe. (2018). A study on legal service of AI. *Journal of The Korea Society of Computer and Information, 23(7),* 105-111.
DOI : 10.9708/JKSCI.2018.23.07.105

[2] Eliot. Lance. (2020). AI and Legal Reasoning Essentials. *LBE Press Publishing.*

[3] Baker. J. J. (2018). 2018: A Legal Research Odyssey: Artificial Intelligence as Disruptor. *Law Library Journal, 110(Issue 1),* 5-30.

[4] Genesereth. M. (2019). Computational law. *Stanford Center for Legal Informatics.*

[5] Markou. C & Deakin. S. (2020). Is Law Computable? From Rule of Law to Legal Singularity. May, 4, 2020. SSRN, *University of Cambridge Faculty of Law Research Paper.*

[6] Aleven. V. (2003). Using Background Knowledge in Case-based Legal Reasoning: a Computational model and an Intelligent Learning Environment. *Artificial Intelligence, 150(1-2),* 183-237.
DOI : 10.1016/S0004-3702(03)00105-X

[7] Hage. J. (2000). Dialectical models in artificial intelligence and law. *Artificial Intelligence and Law, 8(2-3),* 137-172.

[8] Ashley. K., Branting. K, Margolis. H & Sunstein. C. R. (2001). Legal Reasoning and Artificial Intelligence: How Computers "Think" Like Lawyers. *University of Chicago Law School Roundtable, 8(1),* 1-28.

[9] El Ghosh. M. (2018). Automation of legal reasoning and decision based on ontologies. *Doctoral dissertation.* Normandie Universite.

[10] Ho. J. H., Lee. G. G & Lu. M. T. (2020). Exploring the Implementation of a Legal AI Bot for Sustainable Development in Legal Advisory Institutions. *Sustainability, 12(15),* 5991.
DOI : 10.3390/su12155991

[11] Mikolov. T., Chen. K., Corrado. G & Dean. J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781.

[12] Mikolov. T., Sutskever. I., Chen. K., Corrado. G. S & Dean. J. (2013). Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems,* 3111-3119.

[13] Sahlgren. M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies, 20,* 33-53.

[14] Bybee. J. L & Hopper. P. J. (Eds.). (2001). *Frequency and the emergence of linguistic structure (Vol. 45).* John Benjamins Publishing.

[15] Steck. H. (2011, October). Item popularity and recommendation accuracy. *In Proceedings of the fifth ACM conference on Recommender systems,* 125-132.

[16] Bojanowski. P., Grave. E., Joulin. A & Mikolov. T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5,* 135-146.
DOI : 10.1162/tacl_a_00051

[17] Joulin. A., Grave. E., Bojanowski. P & Mikolov. T. (2016). *Bag of tricks for efficient text classification.* arXiv preprint arXiv:1607.01759.

[18] Graves. A & Schmidhuber. J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks, 18(5-6),* 602-610.
DOI : 10.1016/j.neunet.2005.06.042

[19] Gers. F. A., Schraudolph. N. N & Schmidhuber. J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of machine learning research, 3(Aug),* 115-143.

[20] Graves. A., Mohamed. A. R & Hinton. G. (2013, May). Speech recognition with deep recurrent neural networks. *In 2013 IEEE international conference on acoustics, speech and signal processing, IEEE,* 6645-6649.
DOI : 10.1109/ICASSP.2013.6638947

이 영 근(Yeong-Keun Lee) [정회원]

· 1988년 2월 : 연세대학교 경영학과(학사)
· 2017년 8월 : 공주대학교 기계자동차공학과(공학석사)
· 2021년 8월 : 공주대학교 컴퓨터공학과(공학박사)
· 관심분야 : AI, 빅데이터, 영상처리
· E-Mail : johnyklee1@naver.com

박 구 락(Koo-Rack Park) [정회원]

· 1986년 2월 : 중앙대학교 전기공학과(공학사)
· 1988년 2월 : 숭실대학교 전자계산학과(공학석사)
· 2000년 2월 : 경기대학교 전자계산학과(이학박사)
· 1991년 ~ 현재 : 공주대학교 컴퓨터공학부 교수
· 관심분야 : IT 컨버전스, 정보통신, 머신러닝, 전자상거래
· E-Mail : ecgrpark@kongju.ac.kr

이 후 영(Hoo-Young Lee) [정회원]

· 2002년 2월 : 우송대학교 컴퓨터학과(공학사)
· 2017년 2월 : 공주대학교 컴퓨터공학과(공학석사)
· 2020년 2월 : 공주대학교 컴퓨터공학과(공학박사)
· 2020년 3월 ~ 현재 : ㈜이르테크SW 사업부 본부장
· 관심분야 : 빅데이터, 하둡, 정보보안
· E-Mail : hooyoung.paul.lee@gmail.com