**KIPS Trans. Comp. and Comm. Sys.**
Vol.9, No.10 pp.231~238
ISSN: 2287-5891 (Print), ISSN 2734-049X (Online)


https://doi.org/10.3745/KTCCS.2020.9.10.231

# CNN-Based Novelty Detection with Effectively Incorporating Document-Level Information

Seongung Jo† · Heung-Seon Oh†† · Sanghun Im††† · Seonho Kim††††

## ABSTRACT

With a large number of documents appearing on the web, document-level novelty detection has become important since it can reduce the efforts of finding novel documents by discarding documents sharing redundant information already seen. A recent work proposed a convolutional neural network (CNN)-based novelty detection model with significant performance improvements. We observed that it has a restriction of using document-level information in determining novelty but assumed that the document-level information is more important. As a solution, this paper proposed two methods of effectively incorporating document-level information using a CNN-based novelty detection model. Our methods focus on constructing a feature vector of a target document to be classified by extracting relative information between the target document and source documents given as evidence. A series of experiments showed the superiority of our methods on a standard benchmark collection, TAP-DLND 1.0.

Keywords : Deep Learning, CNN, Novelty Detection

# 효과적인 문서 수준의 정보를 이용한 합성곱 신경망 기반의 신규성 탐지

조 성 웅† · 오 흥 선†† · 임 상 훈††† · 김 선 호††††

## 요 약

웹 상에 수 많은 문서가 등장함에 따라 기존 문서와 내용이 중복되는 문서를 찾아서 제외함으로써 새로운 문서를 찾는 노력을 줄일 수 있어 문서 수준의 신규성 탐지(novelty detection)가 중요해졌다. 최근 연구에서는 합성곱 신경망(CNN) 구조 기반의 신규성 탐지 모델 구조가 제안되었고 상당한 성능 향상을 나타내였다. 본 논문에서는 기존의 CNN 기반의 모델에서 문서 수준의 정보가 제한적으로 사용되는 것을 관측하고 문서의 신규성을 결정할 때 문서 수준의 정보가 중요하므로 제한적인 사용이 문제가 된다고 가정하였다. 이에 대한 해결책으로, 본 논문에서는 합성곱 신경망 기반 신규성 탐지 모델 구조를 개선하여 문서 수준 정보를 효과적으로 사용하는 두 가지 방법을 제안한다. 본 논문에서 제안하는 방법은 대상(target) 문서와 증거로 주어진 출처(source) 문서 사이의 상대적(relative) 정보를 추출하여 신규성을 분류할 대상 문서의 특징 벡터를 구성하는 것에 초점을 맞춘다. 본 논문에서는 표준 벤치마크 데이터 셋인 TAP-DLND 1.0를 이용하여 여러 실험을 통해서 제안한 방법의 우수성을 보여준다.

키워드 : 딥 러닝, 합성곱 신경망, 신규성 탐지

## 1. Introduction

Text-level novelty detection is a binary classification task of determining novelty of a given text as novel or non-novel. In common, it is categorized as sentence-level novelty detection and document-level novelty detection according to the scope of a text. Document level novelty detection aims at classifying a document to be novel if it contains sufficient new information compared to documents seen previously. Most previous work in novelty tracks of Text Retrieval Conferences (TREC) from 2002 to 2004 aims at sentence-level novelty detection with relative sentences or documents. In [1] and [2], there are attempts to extend the scope of a text to document-level using classic probability measures. Based on the

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2019R1G1A1003312).
† 비 회 원 : 한국기술교육대학교 컴퓨터공학부 학사과정
†† 정 회 원 : 한국기술교육대학교 컴퓨터공학과 조교수
††† 비 회 원 : 한국기술교육대학교 컴퓨터공학부 석사과정
†††† 비 회 원 : 한국과학기술정보연구원 책임연구원
Manuscript Received : June 3, 2020
Accepted : July 15, 2020
* Corresponding Author : Heung-Seon Oh(ohhs@koreatech.ac.kr)

big success of NLP with deep learning, the recent work [3] proposed a method that uses convolutional neural network (CNN) architecture to document-level novelty detection. The goal of the work is to classify an input target document with three (or more) additional source documents to be novel or non-novel. The key idea of the previous model is to exploit a vector representation capturing relative information, i.e. relative document vector (RDV), between a target document and source documents. It is assumed that the RDV encodes relatively new information of a target document w.r.t source documents. Use of the RDV showed significant improvements compared to existing methods.

In the previous work [3] , a RDV is represented by stacking a set of relative sentence vectors (RSVs). A RSV is constructed as follows. First, every sentence in a target document and source documents is encoded as a vector using a pre-trained sentence encoder. Second, a single most similar source sentence is selected for a target using cosine similarity between two sentence vectors. Third, a RSV is constructed using two sentence vectors with concatenation of results applied by various operators. This top-1 similar sentence selection filters irrelevant information and focuses on most relevant information of source documents. However, we observed that it discards useful document-level information which may be key evidence in determining novelty. We assumed that incorporating more document-level information effectively supports an improvement of novelty detection.

Based on the observation, this paper proposes methods of incorporating document-level information to novelty detection in the context of the previous CNN architecture. The key idea of our methods is to leverage document-level information in the RDV by alleviating the restriction of top-1 sentence selection. Based on the same context of the previous model architecture, our methods are implemented in the RDV generator. A RDV is constructed via relational information selector and relative information aggregator where the former component collects relational information between two sentences by a similarity and the latter component produces relative information with a vector operator. Specifically, the two methods indicate different vector operators using document-level information in the relative information aggregator. A series of experiments show superiority of our methods compared to the previous method in [3]. The contributions of this paper are summarized as follows:

1. We propose methods of exploiting useful document-level information in the context of RDV generation by observing the lack of document-level information in the CNN-based model.
2. We show the superiority of our methods compared to the previous methods by a series of concrete experiments.

The rest of this paper is organized as follows. In Section 2, important related works to novelty detection are introduced briefly. Section 3 describes the details of novelty detection architecture with our methods. Experimental results are delivered with in-depth analysis in Section 4. Finally, we conclude with a summary and our future research direction in Section 5.

## 2. Related Work

Novelty data mining task had originated from the Topic Detection and Tracking(TDT)[4] evaluation task that detects new events or Frist Story Detection (FSD) w.r.t online news stream.

The task grouped mostly involved the news stories into clusters with measuring the belongingness of an input story to any of the clusters based on system.

The novelty detection task has been actively researched since the novelty tracks of Text Retrieval Conferences (TREC) from 2002 to 2004 that was concentrated on sentence-level novelty detection. The goal of these tasks was to capture novelty in each sentence for a given topic and relevant documents.

In 2002, the document level novelty detection has risen to the surface on natural language processing (NLP) area by [1] and [2]. In [1] a topic classifier was proposed for a document that contains novelty and a topic with named entities. A research in [2] has a different view for novelty where five measures are defined to calculates the novelty of an input document using a set of memorized documents. In addition, [2] proposed a method to improve the measures by adding the innovativeness of documents based on information entropy.

Recently, many deep learning-based methods have

been developed for NLP. Text-CNN [5] is the most successful application which adopts CNN to sentence-level sentiment classification. It showed feasibility of adopting CNN architecture to NLP task during recurrent neural network (RNN) architectures are dominant. Since that, CNN becomes a main option to design a model architecture.

CNN model architecture for observing important information in document-level novelty detection is highlighted by [3]. In [3], CNN-based classifier was proposed to detect document novelty. As an input to the classifier, a RDV, simply a set of sentence vectors, is generated by stacking a set of RSVs. A RSV is a vector representation of an sentence produced by bidirectional-long short term memory(Bi-LSTM)[6] based a language model. However, we observed a restriction of document-level information even though it showed the usefulness of CNN model for the document-level novelty detection.

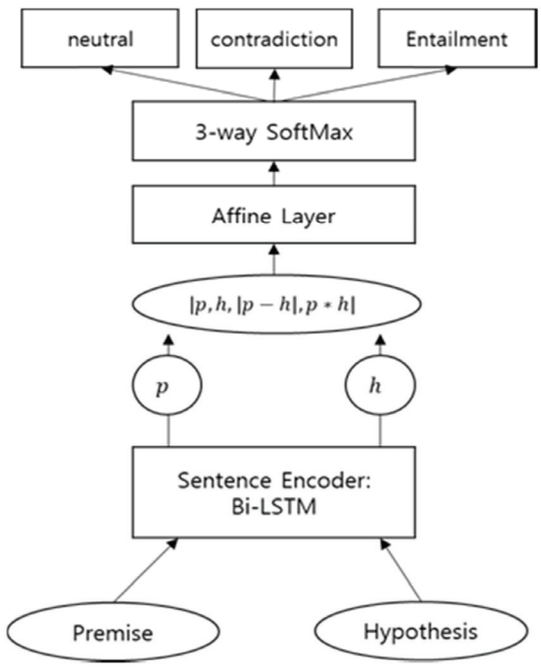



Fig. 1. Overall Architecture of Training Sentence Encoder

## 3. Model

This section describes our model architecture to novelty detection. Our model follows a recent novelty detection research in [3]. The model is composed of two components: sentence encoder and novelty classifier. The sentence encoder generates vector representations of sentences in a document. Using this encoder, sentences of source and target documents are represented to fixed-size vectors and inputted to the novelty classifier. Similar to [6], fine-tuned Bi-LSTM model is utilized as the sentence encoder. The novelty classifier determines novelty of the source document by using the relative information between source and target documents. This relative information is compressed in a RDV. Using the RDV as an input feature vector, a CNN-based binary classifier [5] is trained. We will explain the detail of each component and introduce a new RDV generator which captures better relative information.

### 3.1 Sentence Encoder

The role of a sentence encoder is to represent the context of a given text as a fixed-size vector. Fig. 1 shows the overall architecture of fine-tuning our sentence encoder. Precisely, bidirectional LSTM model (Bi-LSTM) [6] fine-tuned with SNLI dataset was used to our sentence encoder. The aim of fine-tuning is to make the sentence encoder capture universally useful features by learning from premise and hypothesis relations. In [6], the Bi-LSTM sentence encoder was trained to extract two vectors p and h for premise and hypothesis, respectively. Then, a feature vector is generated to retain relative information using equation 1 proposed in [7].

$$[p, h, |p-h|, p^{*}h] \tag{1}$$

Finally, a classifier with the feature vector is trained for sentence judgements. According to this process, we believed that the sentence encoder can be adopted in various NLP tasks without domain-specific adaptation.

### 3.2 Novelty Classifier

This subsection describes the RDV generator and the novelty classifier, respectively. The RDV generator is responsible for creating the relative information between target document and source documents. With the RDV as a feature vector, a CNN-based classifier is trained to determine novelty of a target document.

Fig. 2 depicts a procedure of generating a RDV from a target document $T$ and three source documents $s_i$. For each sentence in $T$ and $s_i$, a sentence vector

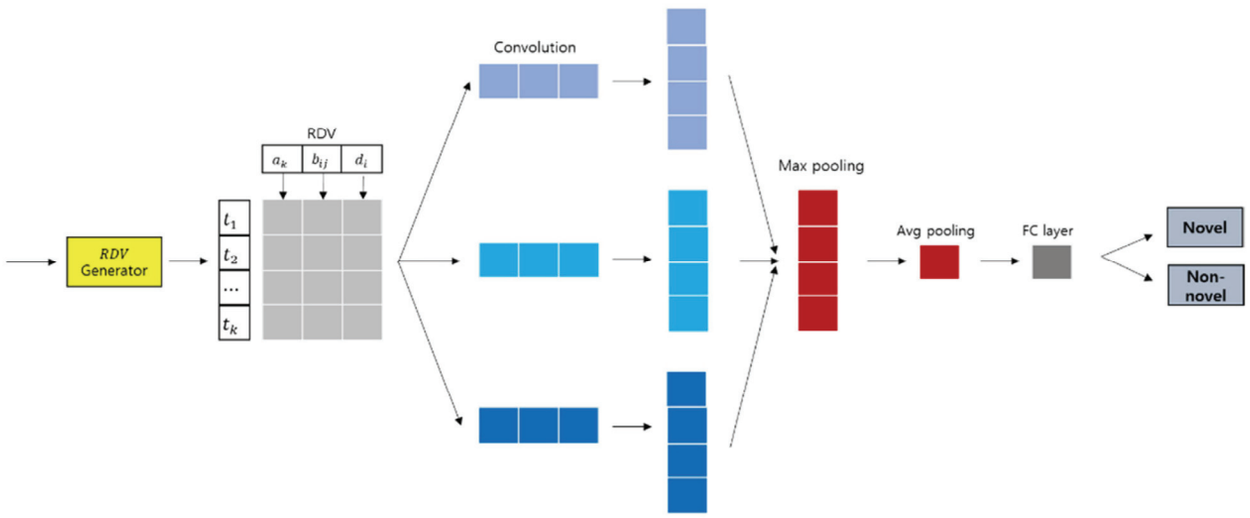



Fig. 2. Classifier Architecture

is generated using the sentence encoder. $a_k$ and $b_{ij}$ denote a vector for k-th sentence in $T$ and for j-th sentence of i-th source document, respectively. The key idea of RDV is how to capture the relative document-level information of a target document compared to source documents. Straightforwardly, it is achieved by two components: relational information selector and relative information aggregator.

1) Relational Information Selector

The role of relational information selector is to find out the relational information that which source document affects a target sentence. It is crucial in determining novelty of a document. Specifically, a target sentence is mapped to a vector generated from a source document selected by a similarity score. The cosine similarity between a target sentence vector $a_k$ and a source sentence vector $b_{ij}$ is computed as:

$$sim(a_k, b_{ij}) = \frac{a_k \cdot b_{ij}}{|a_k||b_{ij}|} \tag{2}$$

For a target sentence vector $a_k$, a source document $c^{*_k}$ with the highest score is selected according to equation 3. Then, a document vector $d_k$ is computed by summing all sentence vectors in a source document $c^{*_k}$ with equation 4.

$$c_k^* = \underset{c \in D_{src}}{\text{argmax}} \left( \sum_{j=1} \text{sim}(a_k, b_{cj}) \right) \tag{3}$$

$$d_k = \sum_{i=c_k^*, j} b_{ij} \tag{4}$$

As a summary, the selector selects the document-level relational information from source documents to a target sentence. This document-level information plays an important role in novelty detection.

2) Relative Information Aggregator

Relative information aggregator aims at building a relative representation of a target sentence by considering the relational information constructed previously. It is defined as a relative sentence vector (RSV). In the previous work [3], three types of operations were applied to construct a RSV to capture the relative information using equation 5. $b_{ij}^{(k)}$ indicates a vector for j-th sentence in i-th source document which has the highest cosine similarity w.r.t a vector for k-th target sentence. As a result, a RSV is a $4 \times D$ sized vector due to concatenation of four vectors resulting from different operations.

$$RSV_k = \left[ a_k, b_{ij}^{(k)}, \left| a_k - b_{ij}^{(k)} \right|, a_k * b_{ij}^{(k)} \right] \in \mathbb{R}^{4D} \tag{5}$$

In equation 5, an RSV has insufficient document-level information since it uses sentence-level information retained in $b_{ij}^{(k)}$ for k-th target sentence. One simple way of incorporating document-level information is to augment $d_k$ using vector concatenation. Definitely,

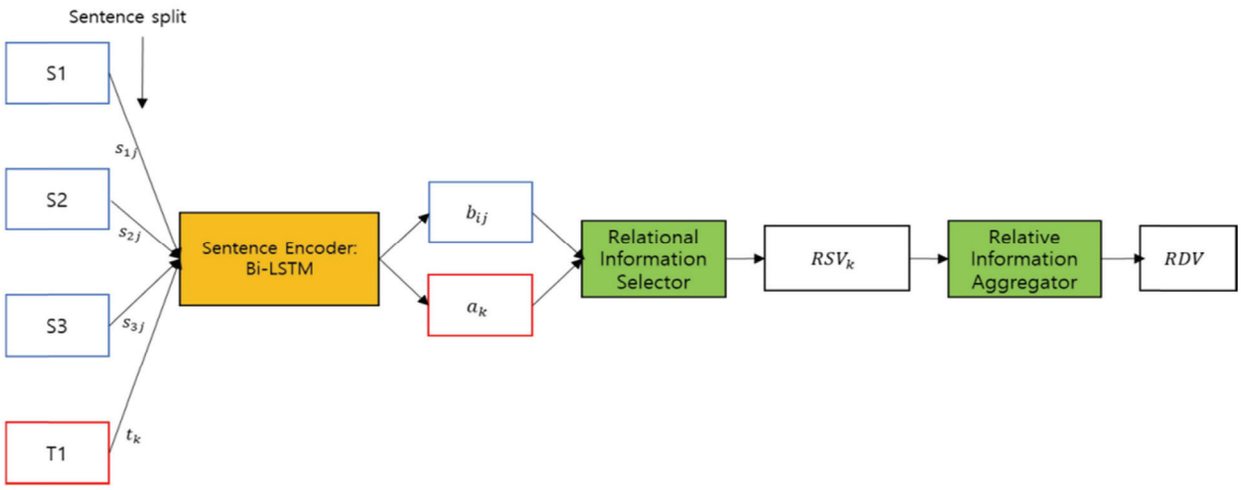



Fig. 3. Procedure of Generating RDV

new RSV in equation 6 conveys more document-level information because $d_k$ is generated by considering all sentence vectors in a source document.

$$RSV_k = \left[a_k, b_{ij}^{(k)}, \left|a_k - b_{ij}^{(k)}\right|, a_k * b_{ij}^{(k)}, d_k\right] \in \mathbb{R}^{5D} \quad (6)$$

We assumed that the method above has a difficulty to properly generate the relative information since it cannot control sentence-level and document-level information due to complex operations, i.e.

subtraction and multiplication, participated in. To avoid this, we simplify the method by removing those operations as shown in equation 7.

$$RSV_k = \left[a_k, b_{ij}^{(k)}, d_k\right] \in \mathbb{R}^{3D} \quad (7)$$

Finally, a target document is converted to an RDV which consists of a set of RSVs. With an RDV as an input feature matrix, a CNN-based novelty classifier is trained.

3) CNN-based classifier

We adopted a CNN model in [5] as our classifier because the effectiveness of the model is shown in various NLP tasks such as [8-10]. As shown in Fig. 3, classifier is composed of convolution, max pooling, average pooling, and fully-connected layer. As the end of the classifier, cross entropy loss is used as our loss function. The detail of all hyper parameters is described in section 4.2.

## 4. Experiments

### 4.1 Data

To directly compare our methods to the previous work, TAP-DLND 1.0 dataset is chosen for evaluation. This dataset [11] is well balanced for classification because it consists of 2,736 novel documents and 2,704 non-novel documents. Each target document is usually mapped to three source documents belonging to 10 social topics such as accident, arts, and crime. Fig. 4 shows the outline of how the dataset is structured.

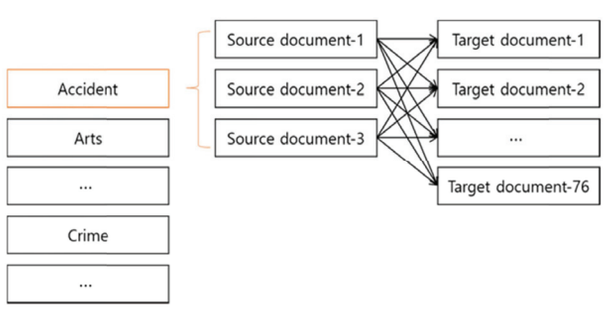



Fig. 4. Structure of TAP-DLND 1.0 Dataset

### 4.2 Experimental Setup

We implemented our model and the baseline CNN model proposed in [3] using PyTorch[1]. To concretely compare our model with the baseline, the same hyper-parameters[2] were set as shown in Table 1. All experiments were performed with NVIDIA GeForce

1) https://pytorch.org/
2) Adam optimizer: parameters set by default variable.

Table 1. Implementation Setup

| Tensor Library | PyTorch |
|---|---|
| Optimizer | Adam [13] |
| Learning rate | 0.0002 |
| Epochs | 250 |
| Input vector size | 2048×3 |
| Number of filters | 100 |
| Filter sizes | [1,3,5] |
| Dropout rate | 0.25 |

RTX 2080Ti * 2. For an evaluation measure, macro F1 is adopted.

### 4.3 Results

We took the results from previous novelty detection models to compare with our methods: Logistic-regression (LR) methods utilizing set difference (SD) , geometric distance (GD) and Dirichlet prior base (LM) proposed in [2], inverted document frequency (IDF) proposed in [12], and CNN-based novelty detection with RDV proposed in [3].

The previous CNN-based novelty detection model is denoted as RDV-CNN[3]. Similarly, we denote that our methods with equations 6 and 7 are RDV-S and RDV-SE, respectively. Table 2 shows the results on TAP-DLND 1.0. As shown, RDV-CNN outperforms the previous linear regression methods (LR-GD, LR-SD, and LR-IDF). It shows the effectiveness of CNN-model in novelty detection. RDV-SE, our method with equation 7, achieved 0.8707 in macro F1 where the performance gain is +2.54 by comparing RDV-CNN. It reveals the importance of relative document-level information.

Table 2. Performance Comparison of Our Model with the Previous Models

| Model | Macro F1 | Gain |
|---|---|---|
| LR with SD | 0.7321 | |
| LR with GD | 0.6984 | |
| LM: Dirichlet Prior + LR | 0.7362 | |
| LR + IDF | 0.5426 | |
| RDV-CNN [3] (Ghosal et al. 2018) | 0.8453 | |
| RDV-SE | 0.8707 | +2.54% |

We further performed experiments to investigate different strategies of exploiting relative document-level information as shown in Table 3. RDV-N indicates RDV-S using equation 6 without the relational information selector. It creates a document vector by summing all source sentence vectors without sentence selection. RDV-NE is based on RDV-N with equation 7. RDV-N and RDV-NE achieved 0.8473(+0.20) and 0.8586(+1.33) compared to RDV-CNN but lower performance compared to RDV-SE. Additionally, we performed experiments to investigate the effect of vector normalization where RDV-NN indicates RDV-N with L2 norm. It achieved 0.8535(+0.82) very close to 0.8533 (+0.82) from RDV-S. In summary, we can conclude that the simple usage of document-level information is the most effective way to novelty detection.

## 5. Conclusion

This paper presents a CNN-based document novelty detection model effectively incorporating document-level information which is crucial in novelty

Table 3. Performance Comparison of Different Strategies based on CNN-Based Novelty Detection, N: Non-selected Model Using All Sentences, S: Selected Model with Eq 4, E: Removing Redundant Embedding with Eq 7, NN: Non-selected Model with L2 Norm

| Model | Description | MacroF1 | Gain |
|---|---|---|---|
| RDV-N | Use all sentence vectors from source docs + eq 6 | 0.8473 | +0.20% |
| RDV-NE | Use all sentence vectors from source docs + eq 7 | 0.8586 | +1.33% |
| RDV-NN | RDV with L2 Norm | 0.8535 | +0.82% |
| RDV-S | Our method with eq 6 | 0.8533 | +0.80% |
| RDV-SE | Out method with eq 7 | 0.8707 | +2.54% |

3) RDV-CNN: Baseline model of our experiments.

classification. In our model, relative information between target and source documents is generated through relational information selector and relative information aggregator in RDV generator. Then, two methods are devised to use document-level information as a feature vector to a novelty classifier. A series of experiments showed superiority of our model and revealed the effectiveness of document-level information in novelty classification. In our future work, we plan to investigate a more advanced way of constructing relative information based on the recent success of graph convolutional networks or graph neural networks and apply our model to more practical domains (e.g. novelty detection in patents).

## References


[1] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, "Topic-conditioned novelty detection," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[2] Z. Yi, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 2002.

[3] T. Ghosal, V. Edithal, A. Ekbal, P. Bhattacharyya, G. Tsatsaronis, and S. S. S. K. Chivukula, "Novelty Goes Deep. A Deep Neural Solution To Document Level Novelty Detection," *Proc. 27th Int. Conf. Comput. Linguist.*, pp. 2802–2813, 2018.

[4] C. L. Wayne, "Topic detection and tracking in English and Chinese," *Proc. 5th Int. Work. Inf. Retr. with Asian Lang. IRAL 2000*, pp. 165–172, 2000.

[5] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 2017-Janua, pp.1746–1751, 2014.

[6] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp.670–680, 2017.

[7] L. Mou et al., "Natural language inference by tree-based convolution and heuristic matching," *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Short Pap.*, pp.130–136, 2016.

[8] B. C. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification," 2014.

[9] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-Aware neural language models," in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, pp. 2741–2749.

[10] A. Conneau, H. Schwenk, Y. Le Cun, and L. Barrault, "Very deep convolutional networks for text classification," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, Vol.1, No.2001, pp.1107–1116, 2017.

[11] T. Ghosal, A. Salam, S. Tiwari, A. Ekbal, and P. Bhattacharyya, "TAP-DLND 1.0: A corpus for document level novelty detection," *Lr. 2018 -11th Int. Conf. Lang. Resour. Eval.*, Vol.7, pp.3541–3547, 2019.

[12] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis, "Using temporal IDF for efficient novelty detection in text streams," pp.1–30, 2014.

[13] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp.1–15, Dec. 2014.



**Seongung Jo**

https://orcid.org/0000-0003-3325-0412

e-mail : oowhat@koreatech.ac.kr

He is an undergraduate student in the School of Computer Engineering at KOREATECH (Korea University of Technology and Education). His research interests are Deep Learning and Natural Language Processing.



**Heung-Seon Oh**

https://orcid.org/0000-0002-9193-8998

e-mail : ohhs@koreatech.ac.kr

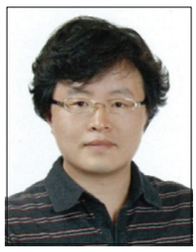
He received a Ph.D. degree in Computer Science from KAIST. in 2014. He has been an assistant professor in the School of Computer Science and Engineering at KOREATECH since 2018. His research interests are in the area of Computer Vision and Natural Language Processing with Deep Learning and Reinforcement Learning.

Sanghun Im

https://orcid.org/0000-0002-2321-9473

e-mail : tkrhkshdqn@koreatech.ac.kr

He is a M.S. course student in the School of Computer Engineering at KOREATECH (Korea University of Technology and Education). His research interests are Deep Learning, Machine Learning and Natural Language Processing.



Seonho Kim

https://orcid.org/0000-0002-9906-878X

e-mail : haebang@kisti.re.kr

Seonho Kim is a principal researcher in the Department of data analysis at the Korea Institute of Science and Technology Information. He received his Ph.D. in Computer Science at the Virginia Tech in 2008. His latest research interests include Machine Learning, Natural Language Processing, and Quantum Computing.