IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 22, 2019, accepted December 30, 2019, date of publication January 1, 2020, date of current version January 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963569

Improved Machine Reading Comprehension Using Data Validation for Weakly Labeled Data

YUNYEONG YANG[1], SANGWOO KANG[2], AND JUNGYUN SEO[1]
[1]Department of Computer Science and Engineering, Sogang University, Seoul 04107, South Korea
[2]Department of Software, Gachon University, Seongnam 13120, South Korea

Corresponding author: Sangwoo Kang (swkang@gachon.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT, under Grant NRF-2019R1C1C1006299.

ABSTRACT Machine reading comprehension (MRC) is a natural language processing task wherein a given question is answered according to a holistic understanding of a given context. Recently, many researchers have shown interest in MRC, for which a considerable number of datasets are being released. Datasets for MRC, which are composed of the context-query-answer triple, are designed to answer a given query by referencing and understanding a readily-available, relevant context text. The TriviaQA dataset is a weakly labeled dataset, because it contains irrelevant context that forms no basis for answering the query. The existing syntactic data cleaning method struggles to deal with the contextual noise this irrelevancy creates. Therefore, a semantic data cleaning method using reasoning processes is necessary. To address this, we propose a new MRC model in which the TriviaQA dataset is validated and trained using a high-quality dataset. The data validation method in our MRC model improves the quality of the training dataset, and the answer extraction model learns with the validated training data, because of our validation method. Our proposed method showed a 4.33% improvement in performance for the TriviaQA Wiki, compared to the existing baseline model. Accordingly, our proposed method can address the limitation of irrelevant context in MRC better than the human supervision.

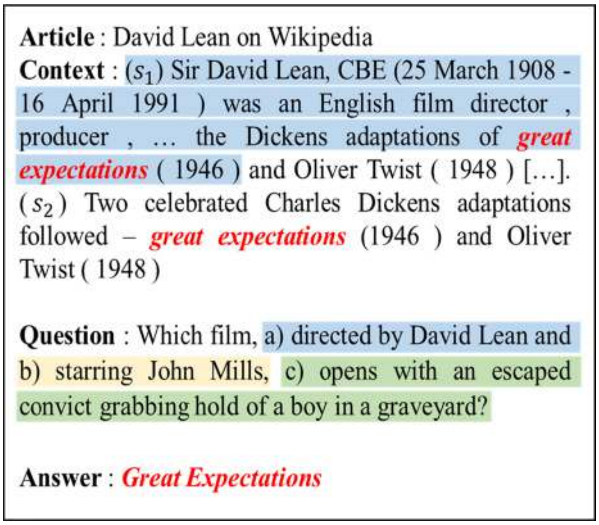
INDEX TERMS Computational and artificial intelligence, data validation, natural language processing, neural networks, machine reading comprehension, weak label.

I. INTRODUCTION

In the past few years, artificial intelligence has seen significant growth in many fields as a result of developments in deep learning [1]–[5]. Natural language processing (NLP), a core technology of artificial intelligence, helps machines to understand, interpret, and manipulate human language. Additionally, because NLP is applicable to all areas in which human language is used, NLP is an extremely crucial task in all domains requiring the use of artificial intelligence.

Therefore, NLP has been actively studied, wherein it has demonstrated sufficient performance in various tasks such as machine reading comprehension (MRC) [6]–[8], machine translation [9]–[11], and natural language inference [12], [13]. MRC, which has recently received a significant amount of attention, is a task wherein an answer is provided to a given query about a text by first understanding the context in which the query arose, i.e. by reading and understanding an entire text pertaining to the query. This process can be considered akin to the task of reading comprehension often used by humans in everyday life; it is necessary for many scenarios such as recommendation systems, question answering, and dialogue. Therefore, machines that use reading comprehension assist people in acquiring information quickly and comfortably. Recently, several approaches [14]–[19] that address the use of large scale datasets for MRC have been proposed; the datasets used in such studies include: Stanford Question Answering Dataset (SQuAD) [20], WikiQA [21], NewsQA [22], and TriviaQA [23]. MRC datasets are composed of context-query-answer triples. Most existing MRC datasets consist of contexts that are well-written and contain sufficient

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.



Article : David Lean on Wikipedia
Context : (s_1) Sir David Lean, CBE (25 March 1908 - 16 April 1991) was an English film director , producer , … the Dickens adaptations of *great expectations* (1946) and Oliver Twist (1948) […]. (s_2) Two celebrated Charles Dickens adaptations followed – *great expectations* (1946) and Oliver Twist (1948)

Question : Which film, a) directed by David Lean and b) starring John Mills, c) opens with an escaped convict grabbing hold of a boy in a graveyard?

Answer : *Great Expectations*

FIGURE 1. Example of irrelevant context.

evidence to answer the query [24]. However, some data from TriviaQA contain a context that lacks sufficient evidence to answer the query. TriviaQA data collects related context for a given query-answer pair, using distributed supervision in Wikipedia or the Web. Furthermore, TriviaQA is a weakly labeled dataset, in which context is collected in a heuristic approach, without human annotation. As an advantage, the TriviaQA data configuration method is a meaningful data organization method with an automatically collected context (for the query); however, as a significant disadvantage, its data is noisy. Such automatically collected data limits improvements to the performance of the model because it includes contexts independent of, and therefore potentially irrelevant to, the query. Because the quality of data directly affects the performance of the model, such mislabeled data has a negative effect on learning.

Figure 1 provides an example of such a query for which irrelevant context is provided. The query asks for the title of the movie which was directed by David Lean, stars John Mills, and starts with a scene where an escaped prisoner holds a boy in the cemetery. The answer to this question is "Great Expectations." To find this answer, three facts must be identified: a) the director; b) the cast; and c) the story. However, only one of these facts is present in the context: in the context we can confirm that David Lean is the director of the movie Great Expectations. However, the context does not include the cast or the story of Great Expectations; thus, the context shown in Figure 1, which does not provide all grounds needed to derive the correct answer, is considered irrelevant context. Removing the data that includes irrelevant contextual information from the training dataset will improve performance, because the data that disturbs the learning process will be removed. In fact, the quality of the datasets used in artificial intelligence is an important issue that must be overcome [25], because a low-quality dataset directly affects their performance. In the Computer Vision field, various studies are being conducted on label noise reduction methods, which can increase the quality of datasets; these include CleanNET [26], DRAE [27], and UOCL [28]. In NLP, extensive research has been conducted on data cleaning processes that enhance the data quality [29], [30]. For instance, data quality has been improved in various ways, such as grammar correction or the removal of stopwords or special characters. Such methods are syntactic data cleaning methods; they process data using rules. However, syntactic data cleaning struggles to deal with contextual noises, such as the irrelevant contextual information in NLP. Moreover, in most of the previous MRC research, the structure of the MRC model was studied in order to increase the performance of the dataset, rather than to solve its underlying problems. Consequently, studies dealing with the problems of the data itself are scarce in the MRC literature.

The TriviaQA includes irrelevant context which is difficult to process using syntactic data cleaning methods, so a semantic data-cleaning method that requires reasoning processes is necessary. Therefore, we propose a new MRC model that utilizes semantic data cleaning.

Our proposed MRC system that involves two steps: a data validation method and a model for finding the correct answer in the refined context. The data validation method removes, from a training set, the data with contexts irrelevant to query resolution; this contributes to enhancing the data quality, as context that does not contain sufficient evidence to answer the query is evaluated as noise and removed from the training set. The answer extraction model is learned using selected data sets through the data validation results; it performs paragraph-selection to process the long text at a paragraph-level. Then, the final answer is extracted using shared normalization for a relative comparison of the correct candidates from several paragraphs. We also evaluate the optimal noise reduction rate from the training data so as to avoid negatively impacting the overall performance of the original task. Accordingly, this work uses deep learning techniques to improve the NLP task of the MRC. Our contributions are as follows.

- We propose a data validation model that removes the irrelevant context within TriviaQA that might impede learning.
- The proposed model does not use syntactic data cleaning techniques. Instead, it uses semantic data cleaning, which verifies data through reasoning processes.
- Experiments confirm that our MRC model outperforms the existing answer extraction model when applied to a TriviaQA verified set without noise.
- The proposed model focuses on adjusting the data itself, as opposed to simply adjusting the structure of the MRC model, with the ultimate objective of providing a more versatile method for an improved MRC performance.

II. RELATED WORK

A. MACHINE READING COMPREHENSION

Among NLP problems, MRC is that which aims to find an answer for a given query according to a context. For existing

Question Answering systems [25] that answer a given query, the answer to the question is found by matching the word or word order contained in the question to those in the sentences of the context text. MRC is different from existing Question Answering systems because MRCs require cognitive processes to understand connotations, such as reasoning using external knowledge, paraphrasing, and multiple sentence reporting [31].

To conduct MRC, data consisting of a triple (question-answer-context) is required. Depending on how the answer is derived, MRC datasets can be divided into three main categories: answer extraction, multiple-choice, and free answering. First, given the context and query, answer extraction asks the machine to extract a span of answers from the context. For this method, which is particularly pervasive in current research, a variety of large-scale benchmark datasets exist, such as SQuAD [20], WikiQA [21], and TriviaQA [23]. Second, using multiple-choice, the right answer is selected from a number of candidates, according to the given context. Third, free answering has no limitations to its answer forms and freely creates the answer to the query. There are several released datasets from which to choose: MS MARCO [33], NarrativeQA [34], MCtest [35], and Race [36].

Among the three types, the answer extraction method has recently become popular with many researchers; it has received such a large amount of much attention that the state-of-the-art is changing frequently. Typical models for implementing MRC include the Bi-Directional Attention Flow [14], the Bidirectional Encoder Representations from Transformers (BERT) [17], and DocQA [18]; a variety of other models are still being proposed [14]–[19].

To find the correct answer, Bi-Directional Attention Flow applies an attention structure to find the context for resolving the query. The BERT model uses unsupervised learning from a large corpus to create a general-purpose language model, fine-tuning it for a specific NLP downstream task through pre-training. A BERT model fine-tuned with a SQuAD dataset has been shown to surpass human performance [17], [37], [38]. Finally, DocQA first ascertains candidate answers in the context at the paragraph-level, and identifies the final answer by comparing the confidence score between the candidates.

The SQuAD outperforms human performance and other datasets because the number of context sentences is small (4-5 sentences); additionally, a simple method of reasoning can be used to find the answer by identifying the most similar sentence for a query that was created by human beings looking at the context text [31]. However, in generating an answer-query for context, TriviaQA is not created by humans. Instead, context is automatically collected using distant supervision for existing answer-query pairs. Because the content for TriviaQA is collected from Wikipedia or the web, the average number of words in the context is 2,895, which is considerably long. Additionally, the queries involved in TriviaQA are more complex than those in SQuAD, and finding the answer requires complex reasoning, such as multi-sentence analysis, rather than finding answers by identifying sentences that are the most similar to the query. Accordingly, DocQA was proposed to address the challenge of TriviaQA's long-length contexts [18]. The DocQA model demonstrated a 10% improvement over the existing model [39], thus solving the problem of TriviaQA through the Paragraph-level QA.

B. BERT

BERT [17] achieved the state-of-the-art through the fine-tuning of the BERT model itself, without the need to attach a new network to handle a particular task. BERT is a language representation model based on the multilayer bidirectional transformer encoder. The use of BERT involves two stages: pre-training and fine-tuning.

First, pre-training is used to build a general-purpose language understanding model that uses unsupervised learning on a large text corpus such as Wikipedia. BERT was simultaneously trained in two tasks: the masked language model and next sentence prediction. In the former, instead of predicting the following word as is done in the existing language model, BERT randomly masks out 15% of the input words and then predicts the masked words. In the latter, when given two sentences, BERT predicts whether the second sentence comes immediately after the first in the corpus. Using these two tasks, BERT constructs a language model, termed the pre-trained BERT.

Second, fine-tuning is conducted as supervised learning which can apply downstream NLP tasks such as MRC [20] and natural language inference [40]. For sentence classification tasks such as natural language inference and semantic analysis [41], the classification (CLS) token, a special token of BERT, is used for fine-tuning.

The first token of every input sequence is a CLS token, which is the special token; the CLS vector of the last hidden layer has the aggregated meaning of the entire sequence representation. Therefore, the CLS vector is used to calculate the probability that a label will be classified. The final hidden state of the CLS token is taken as the fixed-dimensional pooled representation. This is fed into the classification layer, and the label probabilities are computed with a softmax. The parameters of the BERT and the parameters of the classification layer are fine-tuned to maximize the log probability of the correct label such that each task can be performed.

Span-level tasks, such as SQuAD, and token-level tasks, such as Named-entity Recognition [42], only have one more layer than BERT; however, BERT models are fine-tuned similar to the sequence-level. Through two learning methods, pre-training and fine-tuning, BERT obtains new state-of-the-art results on 11 NLP tasks. However, since BERT uses positional embedding, i.e. the method used by Transformer, the maximum number of token inputs is limited to 512.

III. OVERALL ARCHITECTURE

We propose a new MRC model that uses a data validation method to improve the quality of weakly labeled data used

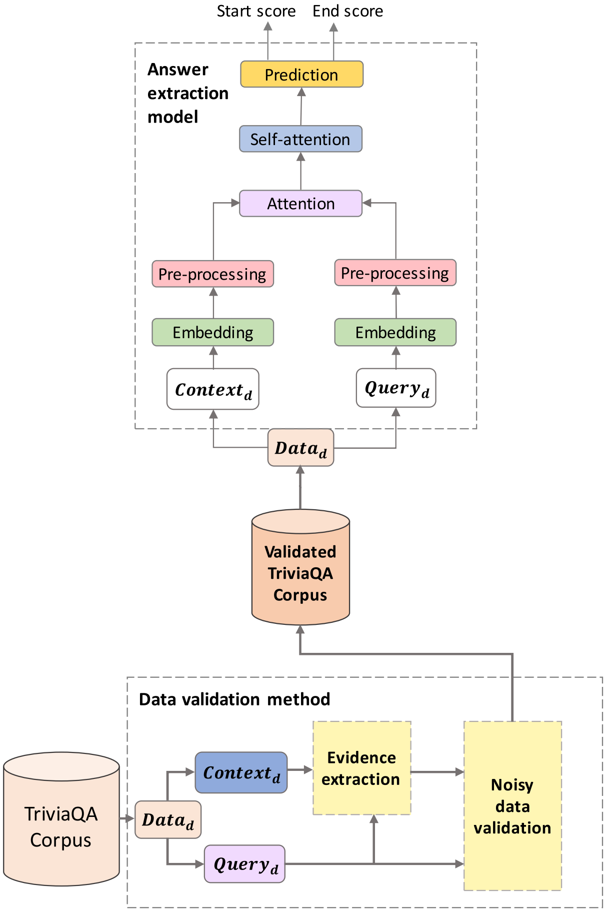


FIGURE 2. **Overall architecture.**

to learn the answer extraction model. TriviaQA [23] is a weakly labeled dataset that automatically collects context on query-answer pairs using distant supervision. Weakly labeled data refers to that which contains context associated with the query, which due to using distant supervision, has insufficient grounds to answer the query; here, such context is termed irrelevant. When irrelevant context is included in the training data, the MRC model is limited in improving its performance.

Here, a novel MRC model is proposed to address this challenge. The novel MRC model proposed herein uses only the data selected through the data validation method to learn the answer extraction model. As shown in Figure 2, the entire process consists of a data validation method and an answer extraction model.

As the first step in the data validation method, evidence extraction selects the paragraphs from the context that relate to the query; then, noisy data validation removes the data consisting of irrelevant evidence with regard to the query. The training data with improved quality, which is obtained through this process, is then used to learn the answer extraction model.

Here, DocQA is used as the answer extraction model. DocQA is a two-step model for dealing with long contexts at a paragraph-level. Paragraphs are selected from the context before the answer candidates are extracted from each paragraph. After comparing the answer candidates extracted from the various paragraphs, the process selects one answer from among the candidates with the highest confidence score.

From the Bi-Directional Attention Flow, BERT, and DocQA models, DocQA was selected as the answer extraction model. In order to monitor the effects of the proposed data validation method, we required an answer extraction model that could process long context effectively. The models that performed well in several MRC datasets were not adequate to process long context; hence, they demonstrated low performance on TriviaQA. Meanwhile, DocQA selects data via paragraph selection to answer queries in advance; thus, it is more appropriate to process long context. We did not use BERT as an answer extraction model though it performs well in MRC tasks, because the input of BERT is restricted to 512 tokens. The word token used in BERT is tokenized by byte pair encoding. Byte pair encoding is an effective tokenizing method to solve Out-of-Vocabulary issues. BERT cannot fully utilize the context of long paragraphs of TriviaQA. If the evidence to answer the query is located in the relevant data after 512 tokens, only the context irrelevant to the query will be used as the input of BERT. In such cases, no evidence for the query will be included in the context. As a result, what is learned with the data appears as if it was learned with irrelevant context. Thus, BERT does not learn to follow the right path to the answer. For the aforementioned reasons, we did not use BERT in answer extraction.

The methodology proposed in this paper is described in Section III.A. In Section III.B, the answer extraction model used in this paper is described in detail.

A. PROPOSED METHOD

In this section, we focus on the two-step method of extraction validation, a key aspect of our proposed model. We propose this method to improve the quality of the training data by removing data with insufficient grounds for the context to be able to sufficiently answer the query. The data validation method consists of two steps: evidence extraction and noisy data validation (Figure 3). Evidence extraction is used to detect the relevance between the query and each paragraph in the context. Resultingly, the paragraph associated with the query becomes the paragraph-level "evidence". Evidence extraction uses BERT to classify the relationship between the query and paragraph. Only extracted evidence is transferred to the next step, i.e. noisy data validation. If no paragraphs are related to the query, the data is removed from the training data in advance.

Noisy data validation is classified using BERT to ensure that the sentence contains sufficient evidence to answer the query. If no sentence within the evidence has a sufficient basis to answer the query, the data containing this context and query is excluded from the training data for the answer extraction model.

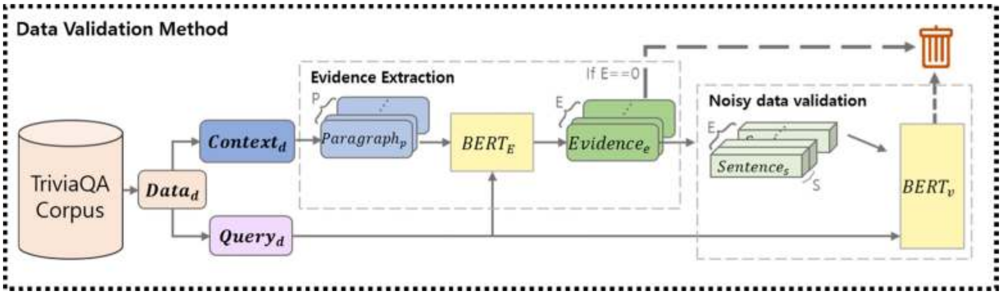


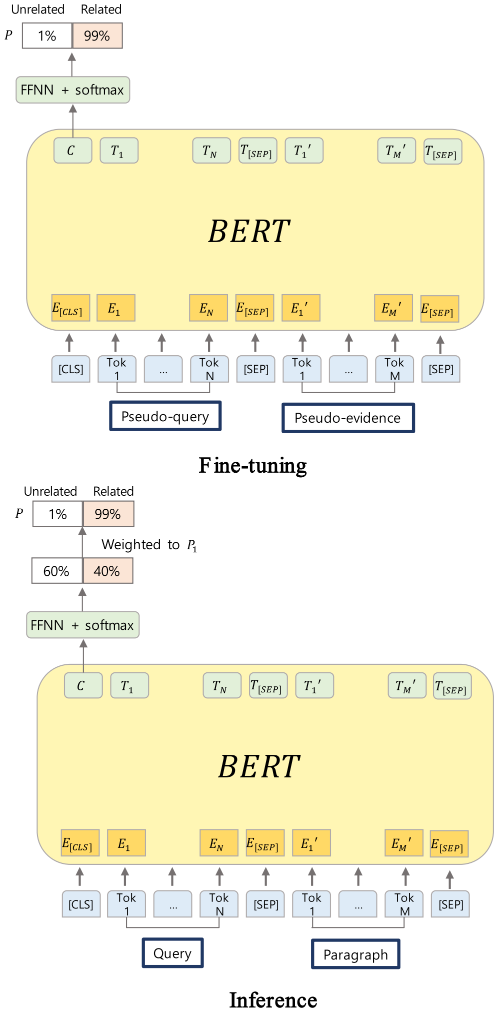
FIGURE 3. Proposed method.

1) EVIDENCE EXTRACTION

The evidence extraction method searches for the paragraphs from the context that are related to the query. To determine paragraph selection, a fine-tuned BERT mode is used for sentence pair classification. Here, the BERT model used for evidence extraction is learned to judge the association between the paragraph and query. To make the association, BERT learns to perform a sentence pair classification task. Sentence pair classification is a task that predicts the relationship between two sentences. For example, Semantic textual similarity deals with determining the extent to which two pieces of texts are similar. Additionally, given a premise, natural language inference is the task of determining whether a hypothesis is entailment, contradiction, or neutral. The input sequence for the sentence pair classification has the form "[CLS] sentence 1 [SEP] sentence [SEP]." As shown in Figure 4(a), the CLS vector in the last hidden layer of the BERT model is used to predict the label. To predict the relevance of the query to the paragraph, BERT learns using the sentence pair classification task from the evidence extraction.

Here, the unsupervised Inverse Cloze Task (ICT) proposed by the Open Retrieval Question Answering System (ORQA) [43] is used to confirm the relevance of the paragraph and query. ICT is a task that finds related context for a sentence, which is the inverse of Cloze task [44]. In the standard Cloze task, the goal is to predict the masked text based on its context. The goal of ICT proposed by ORQA is to find contexts related to a query in a large amount of context.

For the same purpose as ORQA, we construct pseudo-query and pseudo-evidence. Thus, the relationship between query and context in TriviaQA is not learned; instead, it learns the relationship between pseudo-query and pseudo-evidence. Pseudo-query is a sentence, that is not the real query, which is selected at random within the TriviaQA Wiki. Pseudo-query is a declarative sentence; it is different from the actual query, which is an interrogative sentence. ORQA, which uses learned ICT with pseudo-data to predict the context related to the query, performed better than the baseline model [43]. Pseudo-evidence consists of the surrounding sentences of the pseudo-query that are not the context that contains the information about query. If the pseudo-evidence includes the pseudo-query, ICT will learn using word matching. Therefore, 90% of pseudo-evidence is configured not to include pseudo-query.



Fine-tuning

Inference

FIGURE 4. BERT for evidence extraction: (a) fine-tuning BERT for evidence extraction; (b) inference BERT for evidence extraction.

The BERT model uses "[CLS] Pseudo-query [SEP] pseudo-evidence [SEP]" as input for training (Figure 4(a)). The CLS vector, C, which is the last hidden layer representation of the first token in the sequence, is shown in (1). The CLS vector, C, is fine-tuned to learn to predict the relationship between the pseudo-query and the pseudo-evidence. The following methods are used for fine-tuning: (2) produces the probability, P, of predicting the relationship between the pseudo-query and pseudo-evidence, T; as shown in (3), the cross-entropy loss of the predictive probabilities, P, and the relationship between the pseudo-query and pseudo-evidence, T, is computed; all parameters used in the model are learned in such a way that minimizes cross-entropy loss [45].

$$C = \text{BERT}\ (Pseudo\ query, Pseudo\ evidence)[CLS] \in R^H \quad (1)$$

$$P = \text{Softmax}\ (CW^T),\ W \in R^{2*H} \quad (2)$$

$$Cross\ Entropy\ Loss = -\sum_{[0,1]} T_i \log\ (P_i) \quad (3)$$

A fine-tuned BERT model using Pseudo data is used by the evidence extraction to identify the relationship between the query and paragraph of TriviaQA. As shown in Figure 4(b), "[CLS] Query [SEP] Paragraph [SEP]" is used as an input to the BERT model to predict the relevance. We use the final probability P_1 weighted against the probability P_1 as predicted to be relevant. The prediction of relevance by the BERT model, learned with pseudo-data, is represented as 1 (True) if there is no doubt about relevance. The paragraph, which is predicted to be related to the query becomes evidence and moves to the next level; the data that fails to extract any evidence is removed from the training data. If none of the evidence is extracted from one context, the data with that context is immediately removed from the training data.

Here, there are several reasons for extracting the relevant paragraph-level evidence through evidence extraction: First, it is still a difficult problem to express long text information [45]. Therefore, it is not easy to compare the query with a context that contains a lot of information. Second, not all information within the context is related to the query; the information related to the query may only be a part of the context. For example, when the context is Wikipedia's article about AlphaGo, this content contains a variety of information such as history, algorithms, versions, etc. If the query seeks games between AlphaGo and Lee Sedol, the information to answer the query comes from history, not algorithms. Therefore, it is effective to use relevant key information rather than the entire context to determine if the context is related to the query. If information not relevant to query is included, even the best-performing models will not be able to determine it correctly. Thus, we use evidence extraction to find the paragraph level evidence in context, which includes information related to the query.

2) NOISY DATA VALIDATION

Noisy data validation verifies that the evidence extracted (discussed in Section III.A.1) contains the basis for the query. First, it determines whether each sentence in the evidence contains a basis for the query. If no sentence has sufficient grounds for the query, there is not sufficient evidence for the query. We used a fine-tuned BERT model to perform a sentence pair classification to determine whether the sentence includes sufficient grounds for the query.

Here, the BERT model used for noisy data validation learns to determine whether the sentence contains sufficient grounds to answer the query. For this purpose, BERT is learned to perform a sentence pair classification task like evidence extraction. The BERT model, learned from the sentence pair classification task, is used for data validation to verify that the TriviaQA's sentence includes the basis for responding to the query.

The data used for noisy data validation is the Wang dataset [46]. The Wang data was created for the answer sentence selection task used in the traditional Information Retrieval Question Answering (IRQA) [25]. IRQA is conducted in three steps: 1) question processing to analyze the query, 2) paragraph retrieval to find relevant paragraph in the entire document using information retrieval, and 3) answer extraction to find the answer in the paragraph. Of the three stages of IRQA, the answer extraction step should find a sentence that contains the basis for the correct answer among the searched paragraphs. This task is called answer sentence selection. The Wang dataset for answer sentence selection was created using the query and context from the Text REtrieval Conference QA track data [47]. Each sentence in the context is labeled to indicate whether it contains a basis for the query; Wang data is designed to select the appropriate sentence for answering the query from the context.

Therefore, the pre-trained BERT model parameters were fine-tuned using Wang data to perform noisy data validation; this selects the sentence required to answer the query. To train BERT for the objective proposed here, its training input is chosen as "[CLS] query [SEP] sentence [SEP]," as shown in Figure 5(a). The CLS vector, the first of the BERT model's last hidden layer, is used in fine-tuning to determine whether the sentence includes enough basis for query. The fine-tuning method is learned in such a way as to minimize cross-entropy loss of the probability (P), whether or not the sentence contains evidence for query (T), similar to evidence extraction.

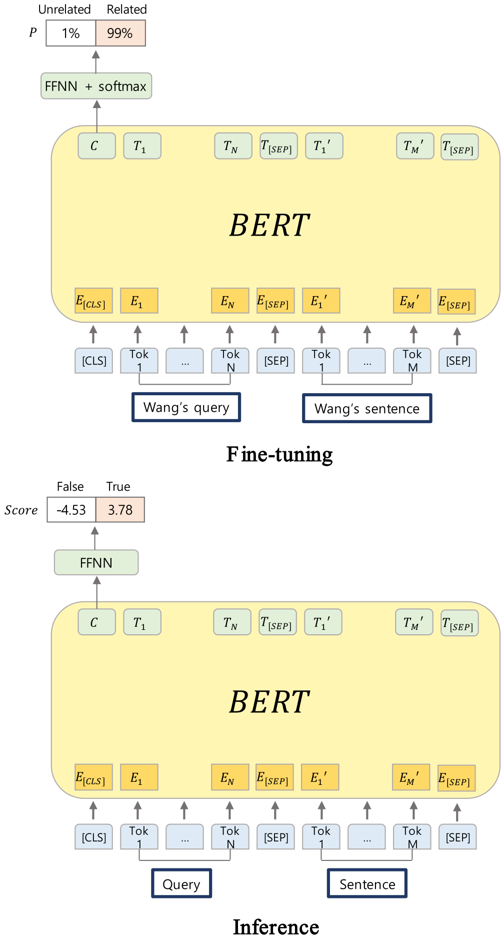
Using the Wang dataset, the fine-tuned BERT model can predict whether the sentence contains the basis for the query to be answered in the TriviaQA dataset. The confidence score is a criterion to determine whether a sentence contains sufficient basis to answer a query. To predict the confidence score of a sentence, the input of the BERT model is "[CLS] Query [SEP] Sentence [SEP]," as shown in Figure 5(b).

The CLS vector (C), which is the last hidden layer representation of the first token in the sequence, is shown in (4). Equation (5) produces the score of the all labels where FFNN refers to a feed forward neural network. Confidence score (CS) is the score for the true label of the score in (6).

$$C = \text{BERT}(Query, Sentence)[CLS] \in R^H \quad (4)$$

$$S = \text{FFNN}(C) \in R^2 \quad (5)$$

$$Confidence\ score = score_{True} \quad (6)$$



Fine-tuning

Inference

FIGURE 5. BERT for noisy data validation: (a) the fine-tuning BERT for noisy data validation; (b) inference BERT for noisy data validation.

The confidence score of evidence has the maximum confidence score among sentences in evidence. If an evidence has a confidence score that does not exceed a certain threshold, it is considered an irrelevant context. Threshold is determined by the train data reduction ratio (K). The data is removed from the training data if all the evidence includes no basis for the query.

B. ANSWER EXTRACTION MODEL

Here, DocQA, a neural question answering model that considers a given context on a paragraph-level, is used for the answer extraction model. DocQA, an answer extraction model, is learned using selected data through the data validation method. DocQA consists of a paragraph selection, answer extraction, and a confidence scoring method. First, paragraph selection selects which paragraphs are used to extract the answer from the entire context. Paragraphs are selected using the Term Frequency and Inverse Document Frequency (TF-IDF) [48] cosine distance of each paragraph and query. If there is one document associated with the query, the paragraph is selected according to the TF-IDF score. In the case of multiple documents, paragraphs are selected using a linear classifier that uses various features and a TF-IDF score.

For the selected paragraphs, the answer candidates for each paragraph are predicted using the paragraph-level answer extraction model for each paragraph input. The answer extraction model predicts the score of the answer span using five layers with context and query: first, in the embedding layer, each word of the context and query is represented using a pre-trained word vector and a character-derived word embedding; second, in the pre-process layer, bi-directional GRU [49] is used to create a word representation with information on bidirectional words of context and query; next, the attention layer uses the Bi-Directional Attention Flow model [14] to create a query-aware context representation; then, the self-attention layer applies self-attention to understand the internal relationship of the context-aware representation created in the previous layer; finally, the prediction layer predicts the start and end score of the answer span using a linear layer and softmax.

DocQA uses a confidence method to handle multiple answer candidates in different paragraphs. The DocQA model predicts the answer for all paragraphs selected in the paragraph selection. The predicted answer is the answer candidate. For each answer candidate, the final answer is the span with the highest sum of the start and end scores. However, each answer score is relative to the words in a paragraph. Thus, the confidence score is local; the global score cannot be used for the comparison of individual paragraphs. Accordingly, the model is learned by modifying the object function in four ways: shared normalization, merge, no-answer option, and sigmoid to compare the scores of different paragraphs in DocQA. The shared normalization method with the best performance is used for the softmax calculation, wherein the performance is normalized for all selected paragraphs within a context. This approach allows a relative comparison between paragraph scores without additional information pertaining to other paragraphs. The paragraph-level question answering model outperforms the previously proposed QA model, to address SQuAD and TriviaQA.

IV. EXPERIMENTS

We conducted experiments to study the performance of our model. The dataset used in the experiments, TriviaQA, is described in section IV.A. In section IV.B, we present results of our proposed model on TriviaQA Wiki.

A. TriviaQA

1) DATASET SPECIFICATION

We evaluated our model on a large-scale reading comprehension dataset, TriviaQA. We experimented with TriviaQA Wiki. Query-answer pairs in TriviaQA were gathered from 14 trivia and quiz-league websites. TriviaQA Wiki contexts were gathered from Wikipedia articles using distributed supervision [23]. Because TriviaQA automatically collected context for query-answer pairs, context did not always

TABLE 1. Error Analysis on TriviQA-wiki.

Category	Ratio (%)
Insufficient evidence	**36**
Correct answer not in the answer key	33
Commonsense knowledge required	4
Multi-sentence reasoning	2
Prediction error	25

TABLE 2. Query-Evidence pairing example.

Question	Which car maker produces the Altea?
Answer	Seat
Context	(1) A video showing new **SEAT**, Škoda & Volkswagen cars being transported by rail at Kutná Hora město train station in the Czech Republic (2) **SEAT** | Spain | Subsidiary | Europe, China, Singapore, Mexico, Central America, South America (except Chile), Middle East, Northern Africa, New Zealand

contain the information necessary to sufficiently answer the query. Therefore, TriviaQA provided a verified set as a test set. In the verified set, humans confirmed in person that the context has the information needed to answer the query. TriviaQA wiki has 110,648 query-context pairs in the training set and 14,229 pairs in the development set. The verified test dataset is composed of 640 pairs that have been made noise-free by human annotators. Therefore, we used this verified dataset for evaluation.

2) ERROR ANALYSIS ON TriviaQA

TriviaQA automatically collected context on query-answer pairs using distant supervision. Thus, TriviaQA contains some irrelevant contexts with insufficient information to answer queries.

We performed an error analysis of the baseline model, DocQA, the result of which motivated our proposed model. Using the DocQA model, we performed error analysis on 100 pieces of sampled data from the development set of TriviaQA Wiki. Resultingly, the errors can be divided into five categories (Table 1). Overall, the occurrence of errors can in part be attributed to the following major reason: although the answer itself is located several times in the context, not every context contains sufficient and obvious evidence from which to draw the correct answer for query resolution.

For example, the query in Table 2 asks for the manufacturer, *SEAT*, which produces the Altea, a car. The context mainly describes the automotive industry, the manufacturer *SEAT* is mentioned in many parts. However, Altea does not appear in the context. Such inconsistency could arise from the limitations in the weak labeling method, that gathers entire contexts by distant supervision, rather by human-annotation. Additionally, the second-largest contributor to the errors is as follows: the context is falsely aliased as a different expression, or the original correct answer has a different meaning. Moreover, there are other minor errors, excluding prediction errors caused by the model itself.

TABLE 3. Hyperparameters for training evidence extraction and noisy data validation.

Hyperparameter	Value
Number of Layers	12
Hidden size	768
Attention heads	12
Learning Rate	5e-5

The overall results of our error analysis show a similar tendency to those in [18]. The research was examined on TriviaQA Web data, wherein a major portion of the errors also result from insufficient evidence. With this limitation, we have proposed a new MRC model which can learn using selected data. The effects of our model are described in the next section.

B. EXPERIMENT SETTING

1) TRAINING DETAILS

In all uses of BERT (i.e. for both evidence extraction and noisy data validation), we initialize from the uncased base model. In Table 3, the hyper-parameters of the fine-tuned BERT commonly used for the evidence extraction and noisy data validation are described. The batch size was 12 and the number of epochs, three, in evidence extraction. In noisy data validation, we used eight for the batch size and three for the number of epochs. The data reduction rate (K) used for the threshold in the noisy data validation method was 20. DocQA was trained with a batch size of 60. The Glove 300-dimensional word vectors were used for word embedding. A dimensionality of 140 GRU and 280 for the linear layers in the DocQA model was used [18].

2) METRICS

There are two types of metrics used to evaluate the MRC models: Exact Matching (EM) and the F_1 score of the words in the answer. The EM is the ratio that represents the extent to which the results predicted by the model and the answer are fully matched. The F_1 score (9) is the harmonic mean of precision, calculated by (7), and recall, calculated by (8). True Positive (TP) represents that the value of the actual class is 'yes,' and the value of the predicted class is also 'yes.' True Negatives (TN) indicate that the value of the actual class is 'no,' and the value of the predicted class is also 'no.' False.

Positives (FP) indicate that actual class is 'no' and the predicted class is 'yes.' False Negatives (FN) indicate that the

TABLE 4. Results on Trivia QA.

Model	Dev		Verified	
	EM (%)	F_1	EM (%)	F_1
Baseline	63.99	68.93	67.98	72.88
Random	61.84	66.43	68.01	71.45
Cosine similarity	61.50	66.84	70.37	76.06
Our	62.02	66.80	**71.72**	**77.21**

actual class is 'yes' but predicted class in 'no.'

$$Precision = TP/(TP + FP) \tag{7}$$

$$Recall = TP/(TP + FN) \tag{8}$$

$$F_1 = 2 \cdot (\text{Precision}\cdot\text{Recall})/(\text{Precision}+\text{Recall}) \tag{9}$$

C. EXPERIMENT RESULTS

1) TriviaQA RESULTS

In Table 4, we show the results of comparisons between the baseline model, random model, cosine similarity model, and our model, in terms of their respective EM and F_1 scores. For the TriviaQA Wiki, the performance of our proposed model registered an EM of 71.72% and an F_1 of 77.21%, for the verified set, which outperformed all other models. We used the same data reduction rate (K) of 20 for other models, except the baseline, to objectively compare of performance. We have thus confirmed that our proposed model demonstrates a higher performance than other models.

In particular, our model has a 4.33% higher F_1 score in the verified set than in the baseline model. The effects of our proposed data validation method could be seen through the random and cosine similarity models. In the case of the Random model, 20% of the model was removed through random sampling in the training dataset. Cosine similarity is a model that calculates the cosine similarity between query and context, and then removes the bottom 20%. To calculate cosine similarity, the representation of query and context applies CLS for BERT, which is fine-tuned with MRPC data. With TriviaQA, we found that the random model had a lower performance than that of the baseline model. The cosine similarity model has 3.18% improvement over the baseline, but it indicates a lower performance than that of our proposed method. For verified sets, in which all contexts contain sufficient evidence to answer the query, we found that the proposed method has an improved performance for TriviaQA (Table 4). However, development sets show a lower performance than that of the baseline because the development set contains irrelevant context, which is noisy data; the training data is also noisy. In fact, we found some cases where the baseline got the correct answer while the proposed model did not among development sets. Such cases were errors caused by the irrelevant context where the context did not include enough evidence to answer to the query as in Figure 1 and Table 2. The reasons why the performance of the development set decreased due to irrelevant context are as follows. If the model is learned from the entire training data containing errors, the answer to the query can be found even if the context does not contain the basis for the query, because the model has learned to predict the answer in the irrelevant context anyway; the model was configured to find the answer to the query while learning the relationship between the unnecessary information and the answer. Therefore, even though the data in the development set contains irrelevant context, the model may still be able to find the answer to the query. However, when the data validation method was used to remove the irrelevant data in the training data, the model was learned by identifying the relationship between the answer words and the basis to answer the query. When validated data is used to learn the model, the irrelevant context does not contain sufficient evidence to answer the query; therefore, the answer to the query is not found in the irrelevant context. In other words, the model that learned with validated data is more ideal than the model that learned with the whole train data.

TABLE 5. Data reduction rate (K).

Rate (K)	Dev		Verified	
	EM (%)	F_1	EM (%)	F_1
0	63.99	68.93	67.98	72.88
5	63.36	68.24	67.34	73.87
10	62.57	67.49	70.71	74.47
15	61.83	66.47	70.71	75.56
20	62.02	66.80	**71.72**	**77.21**
25	62.17	67.11	70.03	75.27

2) EFFECTS OF DATA REDUCTION RATE (K)

The accuracy of the distant supervision is 79.7% for the 986 sampled TriviaQA Wiki data [9]. Moreover, it appears that 20% of the training data has irrelevant contexts that can hinder the improvement of the MRC models. Therefore, we experimented to find the optimal data removal rate (K) that maximizes the performance of the model for the data which contains the answers.

As shown in Table 5, the results of the experiment on the data reduction rate show that the highest performance occurs when 20% of the total training data is removed. As the removal rate for the training data approaches 20%, the performance increasingly improves, compared to the baseline model (Table 5). However, with 25% of training data removed, the performance is lower than the training data with only 20% removed.

3) ABLATION STUDY

We conducted an ablation study on both evidence extraction and noisy data validation by examining the mutual effects (Table 6).

TABLE 6. Ablation study.

	Dev		Verified	
	EM (%)	F_1	EM (%)	F_1
Our method	62.02	66.80	**71.72**	**77.21**
- Noisy data validation	62.26	67.30	69.52	74.22
- Evidence extraction	62.39	67.29	69.70	73.44
- All (baseline)	63.99	68.93	67.98	72.88

The experiment was conducted for the highest performance data reduction rate (K) of 20%. Using both the data validation method and the evidence extraction method helped to improve the performance. We find that extracting paragraphs using evidence extraction is more effective than using the entire long-length context immediately. Resultingly, we have found it effective to select the paragraphs in the context as evidence, and ensure that each piece of evidence contains sufficient grounds to answer the query, rather than to directly identify the relationship the between the query and context. Furthermore, models with validated data show a higher performance than that of the baseline, using all the training data. Therefore, data with a high quality that removes irrelevant data from TriviaQA is effective in learning the answer extraction model.

V. CONCLUSION

Here, we propose a new MRC model that removes the irrelevant context of training data through a data validation method and learns the answer extraction model with improved data quality. To experiment with our new MRC model, we used TriviaQA, which includes irrelevant context, similar to real-world question-answer applications. We found that the performance of TriviaQA, which pairs the answer extraction model with a selection of a data validation method, is superior to the existing baseline model. Results of experiments on the data reduction rate (K) showed a 4.33% performance improvement when 20% of the total training data was removed for the TriviaQA Wiki.

Based on the results of an ablation study, we found that using both steps of data validation helped to improve the performance. The novel MRC model we proposed demonstrated a performance improvement for TriviaQA, showing positive effects on learning the answer extraction model while improving the quality of weakly labeled data. In the future, we intend to extend our work toward a more realistic environment. In particular, our data validation method can be the very steppingstone to intensifying training efficiency over the open-domain resources without any human supervision.

REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[2] V. Kepuska and G. Bohouta, “Next-generation of virtual personal assistants (Microsoft cortana, Apple Siri, Amazon Alexa and Google home),” in *Proc. IEEE 8th Annu. Computing Communication Workshop Conf. (CCWC)*, Jan. 2018, pp. 99–103.

[3] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, “Deep voice: Real-time neural text-to-speech,” in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 195–204.

[4] H. Park, Y. Yoo, Y. Park, C. Lee, H. Lee, I. Kim, and K. Yi, “Toward optimal FPGA implementation of deep convolutional neural networks for handwritten Hangul character recognition,” *J. Comput. Sci. Eng.*, vol. 12, no. 1, pp. 24–35, Mar. 2018.

[5] T. Litman, “Autonomous vehicle implementation predictions: Implications for transport planning,” in *Proc. Transp. Res. Board Annu. Meeting*, 2014, pp. 36–42.

[6] M. Hu, Y. Peng, Z. Huang, X. Qiu, F. Wei, and M. Zhou, “Reinforced mnemonic reader for machine reading comprehension,” in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4099–4106.

[7] R. Yang, J. Zhang, X. Gao, F. Ji, and H. Chen, “Simple and effective text matching with richer alignment features,” in *Proc. 57th Annu. Meeting Assoc. for Comput. Linguistics*, 2019, pp. 4699–4709.

[8] H. Jin, Y. Luo, C. Gao, X. Tang, and P. Yuan, “ComQA: Question answering over knowledge base via semantic matching,” *IEEE Access*, vol. 7, pp. 75235–75246, 2019.

[9] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, “Large language models in machine translation,” in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2007, pp. 858–867.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 5998–6008.

[11] D. Banik, A. Ekbal, and P. Bhattacharyya, “Machine learning based optimized pruning approach for decoding in statistical machine translation,” *IEEE Access*, vol. 7, pp. 1736–1751, 2019.

[12] X. Liu, P. He, W. Chen, and J. Gao, “Multi-task deep neural networks for natural language understanding,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4487–4496.

[13] S. Kim, I. Kang, and N. Kwak, “Semantic sentence matching with densely-connected recurrent and co-attentive information,” in *Proc. AAAI*, vol. 33, Aug. 2019, pp. 6586–6593.

[14] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bidirectional attention flow for machine comprehension,” in *Proc. ICLR*, 2017.

[15] M. Hu, F. Wei, Y. Peng, Z. Huang, N. Yang, and D. Li, “Read+verify: Machine reading comprehension with unanswerable questions,” in *Proc. AAAI*, vol. 33, Aug. 2019, pp. 6529–6537.

[16] Y. Tay, L. A. Tuan, S. C. Hui, and J. Su, “Densely connected attention propagation for reading comprehension,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4906–4917.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[18] C. Clark and M. Gardner, “Simple and Effective Multi-Paragraph Reading Comprehension,” in *Proc. 56th Annu. Meeting Assoc. for Comput. Linguistics*, 2018, pp. 845–855.

[19] S. Back, S. Yu, S. R. Indurthi, J. Kim, and J. Choo, “MemoReader: Large-scale reading comprehension through neural memory controller,” in *Proc. 2018 Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 2131–2140.

[20] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ questions for machine comprehension of text,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2383–2392.

[21] Y. Yang, W. Yih, and C. Meek, “WikiQA: A challenge dataset for open-domain question answering,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2013–2018.

[22] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, “NewsQA: A machine comprehension dataset,” in *Proc. 2nd Workshop Represent. Learn. (NLP)*, Nov. 2016, pp. 191–200.

[23] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, “TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension,” in *Proc. 55th Annu. Meeting Assoc. for Comput. Linguistics*, 2017, pp. 1601–1611.

[24] M. Dunn, L. Sagun, M. Higgins, V. U. Guney, V. Cirik, and K. Cho, ''SearchQA: A new Q&A dataset augmented with context from a search engine,'' Apr. 2017, *arXiv:1704.05179*. [Online]. Available: https://arxiv.org/abs/1704.05179

[25] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé, ''A neural network for factoid question answering over paragraphs,'' in *Proc. EMNLP Conf. Empirical Methods Natural Language Process.*, 2014, pp. 633–644.

[26] K.-H. Lee, X. He, L. Zhang, and L. Yang, ''CleanNet: Transfer learning for scalable image classifier training with label noise,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5447–5456.

[27] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, ''Learning discriminative reconstructions for unsupervised outlier removal,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1511–1519.

[28] W. Liu, G. Hua, and J. R. Smith, ''Unsupervised one-class learning for automatic outlier removal,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3826–3833.

[29] T. Okita, ''Data cleaning for word alignment,'' in *Proc. ACL-IJCNLP Student Res. Workshop*, 2009, pp. 72–80.

[30] E. Rahm and H. H. Do, ''Data cleaning: Problems and current approaches,'' *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, Dec. 2000.

[31] S. Sugawara, K. Inui, S. Sekine, and A. Aizawa, ''What makes reading comprehension questions easier?'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4208–4219.

[32] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, ''On the properties of neural machine translation: Encoder-decoder approaches,'' in *Proc. SSST 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2015, pp. 103–111.

[33] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, ''MS MARCO: A human generated MAchine reading COmprehension dataset,'' in *Proc. ICLR*, 2017.

[34] T. Ko ský, J. Schwarz, P. Blunsom, C. Dyer, K. Hermann, G. Melis, and E. Grefenstette, ''The NarrativeQA reading comprehension challenge,'' *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 317–328, May 2018.

[35] M. Richardson, C. J. C. Burges, and E. Renshaw, ''MCTest: A challenge dataset for the open-domain machine comprehension of text,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 193–203.

[36] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, ''RACE: Large-scale ReAding comprehension dataset from examinations,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 785–794.

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, ''RoBERTa: A robustly optimized BERT pretraining approach,'' Jul. 2019, *arXiv:1907.11692*. [Online]. Available: https://arxiv.org/abs/1907.11692

[38] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, ''Semantics-aware BERT for language understanding,'' Sep. 2019, *arXiv:1909.02209*. [Online]. Available: https://arxiv.org/abs/1909.02209

[39] S. Swayamdipta, A. P. Parikh, and T. Kwiatkowski, ''Multi-mention learning for reading comprehension with neural cascades,'' in *Proc. ICRL*, 2018.

[40] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, ''A large annotated corpus for learning natural language inference,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 632–642.

[41] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, ''SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation,'' in *Proc. 11th Int. Workshop Semantic Eval.*, 2017, pp. 1–14.

[42] E. F. T. K. Sang and F. De Meulder, ''Introduction to the CoNLL–2003 shared task: Language-independent named entity recognition,'' in *Proc. 7th Conf. Natural Lang. Learn. (HLT-NAACL)*, 2003, pp. 142–147.

[43] K. Lee, M.-W. Chang, and K. Toutanova, ''Latent retrieval for weakly supervised open domain question answering,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6086–6096.

[44] N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, and J. Allen, ''LSDSem 2017 shared task: The story cloze test,'' in *Proc. 2nd Workshop Linking Models Lexical, Sentential Discourse-Level Semantics*, 2017, pp. 46–51.

[45] P. T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, ''A tutorial on the cross-entropy method,'' *Ann. Oper. Res.*, vol. 134, pp. 19–67, Feb. 2005.

[46] M. Wang, N. A. Smith, and T. Mitamura, ''What is the Jeopardy model? A quasi-synchronous grammar for QA,'' in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2007, pp. 22–32.

[47] E. M. Voorhees and D. M. Tice, ''Building a question answering test collection,'' in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2000, pp. 200–207.

[48] G. Salton and C. Buckley, ''Term-weighting approaches in automatic text retrieval,'' *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[49] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ''Learning phrase representations using RNN encoder-decoder for statistical machine translation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

YUNYEONG YANG received the bachelor's degree in computer engineering from Kwangwoon University, in 2016. She is currently pursuing the master's degree with the Department of Computer Science and Engineering, Sogang University. She is interested in machine reading comprehension, question answering, sequence labeling, and information extraction.

SANGWOO KANG received the Ph.D. degree in computer science from Sogang University. He was a Research Fellow Professor with Sogang University. He is specialized in natural language processing. He has been an Assistant Professor with the Department of Software, Gachon University, since September 2016. He is currently leading the Intelligent Software and Natural Language Processing Laboratory, Gachon University. He is interested in spoken dialogue interface, information retrieval, text mining, opinion mining, big data, and UI/UX. His recent focus has been in applying deep learning techniques to his research.

JUNGYUN SEO received the B.S. degree in mathematics in 1981, and the M.S. and Ph.D. degrees in computer science from the Department of Computer Science, The University of Texas at Austin, in 1985 and 1990, respectively.

In 1991, he returned to join the Faculty of the Korea Advanced Institute of Science and Technology, Taejon, where he led the Natural Language Processing Laboratory, Computer Science Department. In 1995, he moved to Sogang University, Seoul, and became a Full Professor, in 2001. He serves as the President of the Korea Information Science Society, in 2013. His research interests include multimodal dialogues, statistical methods for NLP, machine translation, and information retrieval.