Automation in Construction 134 (2022) 104061

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

# Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers

Yohan Kim [a], Seongdeok Bang [b], Jiu Sohn [a], Hyoungkwan Kim [a,*]

[a] School of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
[b] AIV company, 338 Gwanggyojungang-ro, Suji-gu, Yongin-si, Gyeonggi-do 16942, Republic of Korea

ARTICLE INFO

*Keywords:*
Bidirectional encoder representations from transformers
Disaster management
Information retrieval
Infrastructure management
Natural language processing
Question answering

ABSTRACT

Manual searching for infrastructure damage information from large amounts of textual data requires considerable time and effort. A fast and accurate collection of damage information from such data is necessary for effective infrastructure planning. In this study, a question answering method was proposed to provide users with infrastructure damage information from textual data automatically. The proposed method relies on a natural language model called bidirectional encoder representations from transformers for information retrieval. From the 143 reports collected from the National Hurricane Center, 533 question-answer pairs were formulated. The proposed model was trained with 435 pairs and tested with the remainder. The model was also tested with 43 question-answer pairs created using earthquake-related textual data and achieved F1-scores of 90.5% and 83.6% for the hurricane and earthquake datasets, respectively.

## 1. Introduction

Substantial disaster-related infrastructure data is required for effective infrastructure management in response to natural disasters. The systematic collection of disaster-related information enables efficient short- and long-term infrastructure investment planning and services. The information is stored in various forms; text format is the most widely used. Analyzing textual data to determine the degree of or potential damage to infrastructure is essential for successful infrastructure management. However, analyzing and retrieving information from news articles, academic papers, or disaster-related local or central government reports is a daunting task due to workforce shortages and budget constraints. Therefore, a method based on a natural language model that automatically processes textual data to retrieve damage information for infrastructure managers is required.

Natural language processing (NLP) allows the automatic retrieval of information from textual data. However, implementation is difficult because of the ambiguity and variability of natural languages. Nevertheless, with the development of NLP techniques based on deep learning, these problems have been resolved to a certain extent [1]. Deep learning-based NLP techniques have made it possible to perform challenging information retrieval tasks efficiently. Specifically, information retrieval through question answering (QA) has attracted considerable attention [2]. Contrary to traditional information retrieval techniques, deep learning-based QA provides users with answers to questions as a piece of specific information not as an entire document [3]. In other words, deep learning-based QA can provide users with more detailed information than traditional information retrieval techniques. Because damages to infrastructure are recorded as a piece of specific textual data, QA can be utilized to obtain this information. For example, given the following textual data, "… The most significant flooding occurred in towns along the Susquehanna River, including Tunkhannock, Pittston, Edwardsville, Nanticoke, Wilkes-Barre, and Harrisburg. In Dauphin and Lebanon Counties in the greater Harrisburg area, nearly 5,000 homes were damaged or destroyed. Numerous roads and 18 bridges were also damaged in Pennsylvania," QA can extract damage information such as "nearly 5,000 homes were damaged or destroyed in Dauphin County, Lebanon County, and Pennsylvania." Thus, infrastructure managers can efficiently extract the desired information from various textual data using QA.

Bidirectional encoder representations from transformers (BERT) is a natural language model that render superior performance in various NLP tasks, including QA [4]. A BERT model performs tasks on the target domain via bidirectional pretraining on language representations and

* Corresponding author.
*E-mail addresses:* homez815@yonsei.ac.kr (Y. Kim), bangseongdeok@gmail.com (S. Bang), jiujohn@yonsei.ac.kr (J. Sohn), hyoungkwan@yonsei.ac.kr (H. Kim).

https://doi.org/10.1016/j.autcon.2021.104061
Received 2 March 2021; Received in revised form 14 October 2021; Accepted 21 November 2021
Available online 2 December 2021

transfer learning to fine-tune the pretrained model parameters. This study proposes a new QA-based method for retrieving infrastructure damage information from textual data using the BERT model, which can learn domain knowledge to identify infrastructure damage. The proposed method involves two steps: 1) retrieval of paragraphs with content related to infrastructure damage and 2) retrieval of specific information regarding infrastructure damage. In the first step, a sentence-BERT (SBERT) model first replaces questions and paragraphs with vector representations, and then the cosine similarity score between the vectors is obtained for each paragraph. Finally, the top 1, 3, 5, and 10 paragraphs with high similarity scores are retrieved. In the second step, a BERT model extracts specific information on infrastructure damage from the paragraphs retrieved in the first step. The BERT model was trained with a dataset comprising 533 question–answer pairs collected from 143 tropical cyclone reports published by the National Hurricane Center (NHC) [5]. Of the 533 pairs, 435 were used for training and 98 for testing. In addition, 43 question–answer pairs, created from earthquake-related textual data (three Wikipedia articles and five reports) [6–10], were used to test the proposed model. The experimental results show that the proposed method can effectively retrieve infrastructure damage information from textual data.

## 2. Literature review

### 2.1. NLP for architecture, engineering, construction, facility management, and disaster management

NLP has been used to process and analyze textual data, such as construction and disaster-related documents, in architecture, engineering, construction, facility management, and disaster management. Some studies have used NLP to analyze clauses and extract information from construction project contracts. Al Qady and Kandil [11] suggested concept relation identification using shallow parsing to extract semantic knowledge from construction contract documents. Lee et al. [12] developed a rule-based NLP model for extracting poisonous clauses from construction contracts. Hassan and Le [13] developed a classification model that can automatically identify the requirements in construction contracts to reduce reading time and improve contract scope comprehension. Lee et al. [14] proposed a risk assessment model for identifying contractor-friendly clauses missing from the modified contracts of owners from the contractors' perspective.

Some researchers have used NLP to analyze construction accident reports, such as those from the Occupational Safety and Health Administration, to automatically identify the type and cause of accidents and extent of damage. Kim and Chi [15] proposed a system that automatically retrieves accident cases and extracts tacit knowledge, such as hazardous objects and positions, work processes, and results of accident cases. Zhang et al. [16] used NLP and machine learning to classify construction accident reports according to accident type and identify hazardous objects that could cause accidents. Baker et al. [17] applied NLP and deep learning to extract injury precursors, such as incident type, injury type, body part, and severity from construction injury reports. Fang et al. [18] proposed a BERT-based method to classify information from safety reports automatically.

NLP has also been employed to support automated compliance checking (ACC) during construction. Salama and El-Gohary [19] presented a new approach for ACC of construction operation plans using deontology, deontic logic, and NLP. They also proposed a semantic and machine learning-based method for classifying clauses and subclauses [20]. Zhang and El-Gohary [21] proposed an NLP-based method to extract regulatory requirements automatically from construction regulatory textual documents and transform them into a formalized format for automated reasoning. They also employed a semantic and rule-based NLP approach to derive information automatically from construction regulatory documents [22]. Zhou and El-Gohary [23] proposed a machine learning-based NLP algorithm to categorize clauses in environmental regulatory documents for automated environmental compliance checking. Zhang and El-Gohary [24] proposed a new ACC system integrating semantic NLP and logic-based representations to extract and transform information from textual regulatory documents and automated reasoning. Xue and Zhang [25] presented an NLP-based part-of-speech tagging to extract and convert regulatory information from building codes into computable representations. Mo et al. [26] developed a machine learning model using NLP that automatically assigns personnel to improve the productivity of building staff assignments. Fan and Li [27] introduced a new method for retrieving similar historical cases from a case library using text-mining techniques to avert construction accidents.

Many studies have focused on analyzing and extracting disaster-related textual information from various sources, such as social media and news articles, for disaster management. Ragini et al. [28] categorized disaster-related tweets into five classes (water, food, shelter, medical emergency, and electricity) and used sentiment analysis to classify whether or not the person who posted the tweet required assistance. Wang and Taylor [29] demonstrated the negative correlation between the intensity of an earthquake and human sentiment by applying sentiment analysis to earthquake-related tweets. Yu et al. [30] used NLP and convolutional neural network (CNN) to categorize hurricane-related tweets into five classes (caution and advice, casualties and damage, information sources, infrastructure and resources, and donation and aid). Hao and Wang [31] relied on image and textual data to assess disaster severity and damage type, respectively. Wang and Taylor [32] suggested an emergency detection method for urban areas using topic modeling, an NLP technique that automatically clusters textual data by topic to geotagged tweets. Kundu et al. [33] classified disaster-related tweets into seven classes using long short-term memory (LSTM). Chowdhury et al. [34] proposed a Bi-LSTM model to extract key phrases from disaster-related tweets. Wang and Stewart [35] extracted spatiotemporal and semantic information of hazard events from hazard-related news reports using NLP and hazard ontology.

Several studies have also sought to develop systems for bidirectional communication between users or between users and computers during disasters by analyzing questions and documents based on keywords. The disaster management system by Zheng et al. [36] allowed private and public participants to share and collaborate on disaster information. Chan and Tsai [37] proposed a QA dialogue system that provides information for emergency operations.

The retrieval of disaster information from textual data is challenging because of the complexities in comprehending natural language. Hence, previous studies have used deep learning models, such as CNN and recurrent neural network (RNN), focusing on keywords related to disasters in textual data. CNN-based models can accomplish tasks, such as sentiment analysis and text classification but do not reflect sequential information and are unsuitable for tasks requiring an understanding of the entire text. Although RNN-based models can reflect sequential information, they have a long-term dependency problem, and performance decreases as the length of the input sequence increases [38]. Therefore, a fast and accurate QA disaster management model is

**Table 1**
Summary of NLP-based studies for construction and disaster management (machine learning: ML and deep learning: DL).

| Reference | Task | Algorithm |
|---|---|---|
| Al Qady and Kandil [11] | Concept relation extraction from construction documents | Rule-based |
| Lee et al. [12] | Poisonous clauses extraction from construction contracts | Rule-based |
| Hassan and Le [13] | Requirements identification from construction contracts | ML-based |
| Lee et al. [14] | Missing contract conditions identification from construction contracts | Rule-based |
| Kim and Chi [15] | Construction accident case retrieval and analyses | Rule-based and ML-based |
| Zhang et al. [16] | Construction site accident analysis | Rule-based and ML-based |
| Baker et al. [17] | Construction injury precursors analysis | ML-based and DL-based |
| Fang et al. [18] | Text classification of near-miss information from safety reports | DL-based |
| Salama and El-Gohary [19] | Automated compliance checking of construction plans | Rule-based and ML-based |
| Salama and El-Gohary [20] | Semantic text classification for automated compliance checking | ML-based |
| Zhang and El-Gohary [21] | Information transformation for automated compliance checking | Rule-based and ML-based |
| Zhang and El-Gohary [22] | Semantic information extraction from construction regulatory documents | Rule-based |
| Zhou and El-Gohary [23] | Domain-specific hierarchical text classification for automated compliance checking | ML-based |
| Zhang and El-Gohary [24] | Information extraction and transformation from regulatory documents using a unified automated compliance checking system | Rule-based |
| Xue and Zhang [25] | Building codes part-of-speech tagging for automated compliance checking | Rule-based and ML-based |
| Mo et al. [26] | Staff assignment for building maintenance | ML-based |
| Fan and Li [27] | Retrieving similar cases for alternative dispute resolutions | Rule-based |
| Ragini et al. [28] | Emergency text classification from social media using sentiment analysis | ML-based |
| Wang and Taylor [29] | Coupling sentiment and human mobility in natural disasters using social media | ML-based |
| Yu et al. [30] | Real-time social media text classification for situation awareness | DL-based |
| Hao and Wang [31] | Rapid disaster assessment using multimodal social media | ML-based |
| Wang and Taylor [32] | Urban emergency detection from social media using topic modeling | ML-based |
| Kundu et al. [33] | Classification of short-texts generated during disasters | DL-based |
| Chowdhury et al. [34] | Keyphrase extraction from disaster-related tweets | DL-based |
| Wang and Stewart [35] | Spatiotemporal and semantic information extraction from news reports | Rule-based |
| Zheng et al. [36] | Information sharing and collaboration for efficient disaster management | ML-based |
| Chan and Tsai [37] | Developing question-answering dialogue system for emergency operations in disaster management | ML-based |

necessary.

### 2.2. BERT

Deep learning allows the high-performance execution of many NLP tasks. RNN-based sequence-to-sequence models, including LSTM [39] and gated recurrent unit [40], have been widely used to process sequential data such as text. To overcome the long-term dependency problem in RNN models, an attention mechanism that allows the dependency to be modeled, regardless of the distance of the input or output sequence, was applied to sequence-to-sequence NLP models [41]. The transformer, which is an encoder–decoder model based entirely on the attention mechanism, has improved the performance of many NLP tasks [42]. The BERT model is a language model composed of transformer encoder blocks. Language models assign probabilities to word sequences or sentences, indicating the appropriateness of combining these words or sentences.

The basic concept of the attention mechanism involves focusing on input words that are more relevant to predicted words (output words). Each input word has three hidden states: query (Q), key (K), and value (V). An attention score is first calculated using the dot product of Q and K. The dot product is a pairwise multiplication between two embedded vectors. The attention score is then normalized to the attention distribution by a Softmax activation function. Each value in the attention distribution represents the weight of words, which is a value between 0 and 1. The weight represents the relationship between two words, that is, the larger the weight, the stronger the association. Finally, the attention value is derived using the weighted sum of the values (V). The output of the Softmax function is multiplied by V, and the results are summed to obtain the attention value. Transformer-based models have multi-head attention that computes and concatenates sets of attention values and can capture more information than a single attention head. Attention is calculated using the scaled dot product in BERT, as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q, K, V, and $d_k$ denote the queries, keys, values, and dimensions of the keys, respectively. For large $d_k$ values, $\frac{1}{\sqrt{d_k}}$ functions as a scaling factor to prevent the gradients from becoming extremely small as the dot product increases [42].

To train the model in unsupervised learning, BERT uses a large corpus (BooksCorpus and Wikipedia) to build a general-purpose language model. It also uses domain knowledge to perform downstream NLP tasks in supervised learning. BERT is pretrained using the masked language model (MLM) and next sentence prediction (NSP). The MLM task randomly replaces 15% of the input words with "masked" tokens and predicts them; 80% and 10% of the masked tokens are replaced by [MASK] and random tokens, respectively, whereas 10% remains unchanged. The BERT model enables contextualized word embedding by learning deep bidirectional representations through the MLM task. Word embedding (or token embedding) is mapping words or tokens into vectors for computers to understand natural language. The NSP task determines whether the two sentences are naturally associated. BERT then learns the relationship between sentences and improves the performance of QA and natural language inference tasks. Depending on its size, several BERT models are available, including BERT-Base (12 layers, 768 hidden sizes, 12 attention heads, and 110 million parameters) and BERT-Large (24 layers, 1024 hidden sizes, 16 attention heads, and 340 million parameters). BERT-Large is employed to retrieve infrastructure damage information because it generally performs better than BERT-Base in various NLP tasks, including QA [4].

The SBERT model is a modified BERT model that uses a Siamese network structure. It exhibits excellent performance in several tasks requiring the ability to identify text relatedness (e.g., large-scale semantic similarity comparison, clustering, and information retrieval via

**Table 2**
Previous studies using deep learning-based method for construction and disaster management.

| Model | Reference | Task | Representation | Data source |
|---|---|---|---|---|
| CNN | Yu et al. [30] | Text classification for situation awareness | Word2Vec | Twitter |
| RNN | Kundu et al. [33] | Text classification for disaster management | Word2Vec | Twitter |
| CNN + RNN | Baker et al. [17] | Text classification for construction injury precursors | Word2Vec | Safety reports |
| | Chowdhury et al. [34] | Keyphrase extraction for disaster management | ELMo, IPA, POS | Twitter |
| BERT | Fang et al. [18] | Text classification of near-miss information | Token, Positional, Segment | Safety reports |

semantic search) [43]. For one-shot image recognition, Siamese neural networks consisting of two CNNs that share weights to determine whether two images belong to the same class based on the similarity probability of their vector representations have been proposed [44]. The SBERT model has a Siamese network structure comprising a BERT network instead of a CNN and text as input instead of images. The SBERT model adds a pooling layer after each BERT network to derive fixed-sized embedding. A standard method for using BERT to represent text (e.g., sentences and paragraphs) as a single vector is averaging the vector representation of words in the text or using the vector representation of the first token ([CLS] token) [4,43]. However, because this vector representation does not accurately reflect the meaning of the text, it performs poorly in several tasks, such as clustering and information retrieval [43]. SBERT is designed using the Siamese network structure to resolve this problem, allowing the meaning of the text to be sufficiently reflected in a single vector representation [43]. The pretrained SBERT satisfactorily performs semantic textual similarity tasks without fine-tuning using the domain dataset [43]. Accordingly, the pretrained SBERT model was used in this study to retrieve paragraphs containing infrastructure damage information.

Prior to BERT, deep learning language models were either unidirectional (left-to-right or right-to-left) or shallow bidirectional language models [4]. In contrast, BERT can perform bidirectional deep language learning, thus enabling a contextual understanding of texts and improved performance in QA tasks [4].

Previous studies using construction and disaster-related textual data are presented in Table 1. These studies are divided into rule-based, machine learning-based, and deep learning-based methods. Table 2 shows the models and vector representation methods used in deep learning-based studies [17,18,30,33,34]. CNN and RNN based studies used pretrained embedding models such as Word2Vec and Embeddings from Language Model (ELMo) for vector representation of tokens. Models that can better reflect contextual information, such as ELMo and BERT, have been recently developed. Chowdhury et al. [34] used three types of embedding information: contextual information using ELMo, phonetic and phonological identification information using the International Phonetic Alphabet (IPA), and phonological embeddings, and part-of-speech (POS) tagging information. Through experiments, they demonstrated that additional information improved the performance of the model. BERT uses wordpiece tokenization for token embedding [4], dividing words into common subword units called “wordpieces.” This tokenization method allows rare words to be handled effectively; tokenization of unknown words is possible without special processing [45]. For example, the word “Imelda” that is not in BERT’s vocabulary is tokenized as (‘im’ ‘##eld’ ‘##a’). Because infrastructure and disaster-related textual data contain many domain-specific rare words, BERT can outperform traditional models. The attention layer in BERT is more efficient than the convolutional and recurrent layers and has lower computational complexity, including cost. Because the attention layer can be parallelized, it has a higher computational speed than the recurrent layer. In addition, because of equal attention between all positions, the path length of long-range dependencies within a network is reduced [42]. Hence, BERT can process long texts, such as reports, more efficiently than CNN and RNN. However, BERT has limitations in retrieving infrastructure damage information for disaster-related questions from the given textual data. These limitations are attributed to a lack of relevant domain knowledge. Therefore, textual data related to disasters and infrastructure should be collected, and a QA dataset should be created to train the BERT further to resolve these problems. Accordingly, this study proposes a novel BERT-based QA method to provide users with infrastructure damage information derived from disaster-related textual data.

## 3. Methodology

As shown in Fig. 1, the proposed method uses SBERT and BERT models to retrieve infrastructure damage information from textual data. The pretrained SBERT model retrieves paragraphs containing infrastructure damage information, and the fine-tuned BERT model retrieves the information best suited to the question from the paragraphs using a domain dataset. Since the first step is to calculate the cosine similarity between the question and paragraph, each question and paragraph should be represented as a single vector of size 1xN. SBERT excels at representing a sentence or paragraph as a single vector and outperforms BERT in related tasks [43]. However, since SBERT was designed to efficiently perform tasks that require sentence-level operations, it is not suitable for tasks that require token-level operations such as QA task. Therefore, SBERT was used in the first step, and BERT was used in the second step.

### 3.1. Building NHC dataset to train BERT for QA

NHC reports were used to generate a dataset in the disaster and infrastructure domain to fine-tune the BERT. The paragraphs containing infrastructure damage information were extracted from the reports, and one or more question–answer pairs were established for each paragraph. Because disasters can occur in multiple areas, damage information is provided separately for each area. Accordingly, the name of the disaster (e.g., Hurricane Harvey) and area (e.g., Darlington County) were included in the questions to allow the model to accurately provide information regarding the damage in each area. The answers are provided in phrases or sentences and include the type and extent of damages and

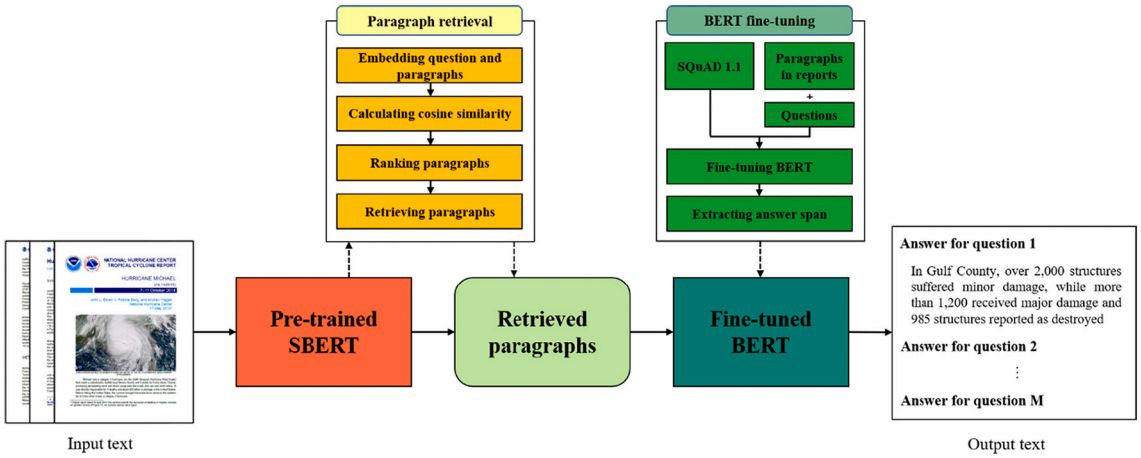


**Fig. 1.** Overview of BERT-based method for retrieving infrastructure damage information.

the infrastructure damaged. We created 533 question–answer pairs 143 reports, of which 435 pairs from 112 randomly selected reports were used to train the BERT, and 98 pairs from the remaining 31 reports were used for model testing.

### 3.2. Retrieval of paragraphs with content related to infrastructure damage using SBERT

The process of retrieving paragraphs related to infrastructure damage is illustrated in Fig. 2. SBERT model comprising two DistilBERT networks [46] pretrained with the Microsoft machine reading comprehension (MS MARCO) dataset [47] was used. The DistilBERT network is a compact BERT model trained through knowledge distillation—a compression technique that transfers knowledge from a large or an ensemble model to a smaller model. The model removes token-type embeddings (segment embeddings) and poolers (the last hidden state of a [CLS] token that is used for the next sentence classification task) from a BERT-based model and reduces the total number of layers by half. The DistilBERT network is 60% faster than BERT-Base and retains 97% of its language comprehension capabilities, even with a 40% reduction in the BERT-based model size [46]. MS MARCO is a large-scale dataset typically used for a variety of tasks, such as passage ranking. The dataset comprises 1,010,916 questions, 8,841,823 passages, and 3,563,535 web documents collected from Bing.

The questions and paragraphs were converted into vectors by the pretrained SBERT model, and the similarity score between each question and paragraph was calculated using cosine similarity. The paragraphs were ranked according to the highest similarity score, and the top K (where K = 1, 3, 5, and 10) paragraphs were retrieved and used for open-domain QA [48–50]. The maximum value of K was set to 10 in this study, considering the number of paragraphs in these reports.

### 3.3. Retrieval of specific information regarding infrastructure damage using BERT

As shown in Fig. 3, this study uses a BERT-Large model with 24 layers. The model input consists of a [CLS] token to represent the start, [SEP] tokens to indicate the end of the question and paragraph, [PAD] tokens to equalize the input length, and general tokens to represent contents. In this study, the maximum input length (maximum number of input tokens) of the BERT model is 512; in texts exceeding this limit, only 512 tokens are input into the model. The input embeddings are

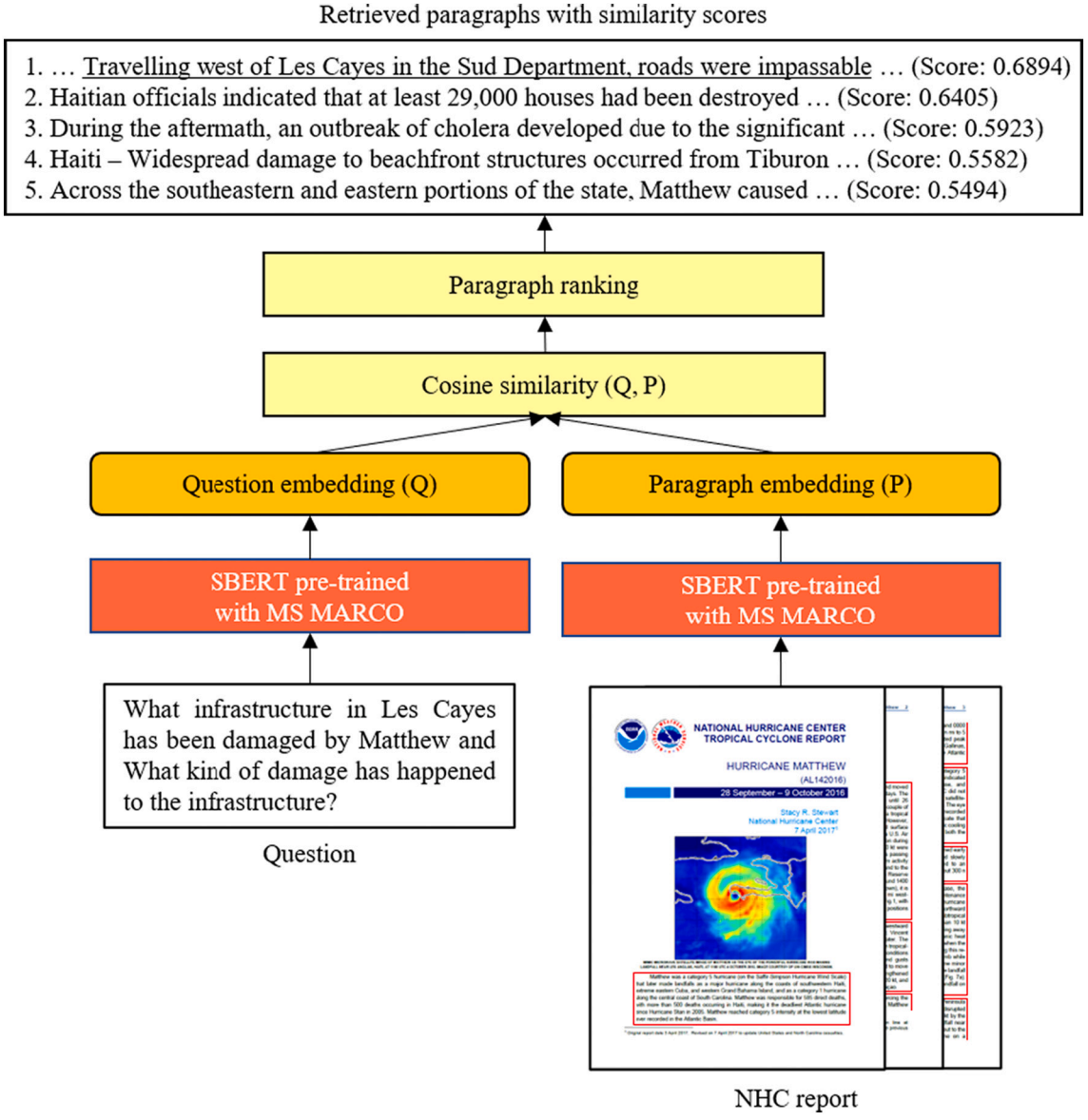


**Fig. 2.** Process of retrieving paragraphs related to infrastructure damage.

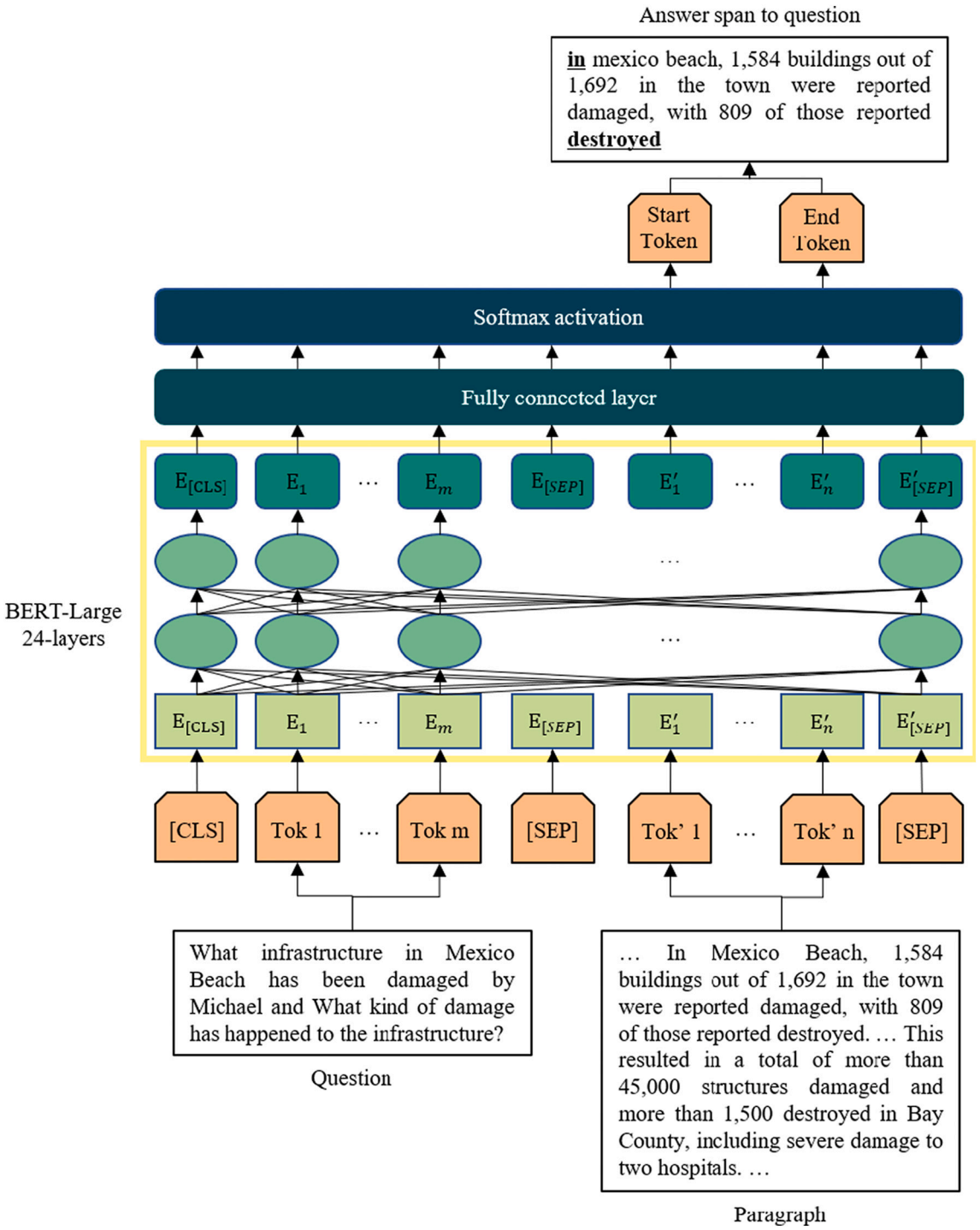


**Fig. 3.** Fine-tuning BERT model for QA.

obtained as the sum of the token, segment, and position embeddings. Token embeddings are vector representations that reflect word features, and each token has a different value. Segment embeddings have a value of 0 or 1, which separates the paragraph from the question. Position embeddings lie between −1 and 1 and provide sequential information to each token because the tokens are not input in order. Input embeddings are adjusted by passing through 24 layers and one fully connected layer. The two probability distributions and input embedding loss values are then obtained via the Softmax activation and cross-entropy loss functions, respectively. The Softmax function represents the probability that the token is the start or end of the answer. The token with the highest probability is selected as the start token. The token with the highest probability among the tokens located behind the start is selected as the end token. All tokens in between are extracted as the answer span to retrieve infrastructure damage information from the paragraphs of disaster reports.

## 4. Experiment

### 4.1. Experimental environment and data description

The proposed method was implemented using Python on a computer

**Table 3**
Summary of the NHC and earthquake datasets.

| Type of disaster | Type of infrastructure | Damage description |
|---|---|---|
| Tropical storm / Hurricane (143 reports from 1998 to 2020)<br>Earthquake (2004 Indian ocean, 2010 Chile, 2010 Haiti, 2011 Tohoku, 2015 Nepal) | Airport / Platform / Railroad / Station / Subway<br>Apartment / Building / Hotel / Resort<br>Construction / Infrastructure / Facility / Structure<br>Aquarium / Casino / Museum / Prison<br>Barn / Farm / Garage / Warehouse<br>Bridge / Pier / Viaduct<br>Business / Church / Office / School<br>Dam / Dock / Levee / Port / Sea wall<br>Highway / Interstate / Road / Roadway / Street<br>Home / House / Residence / Roof<br>Hospital / Medical Center • Facility<br>Pipeline / Power line • pole / Utility<br>Plant / Tower / Tunnel<br>Restaurant / Shop / Store | Blow out / Breach / Broken / Crack / Fail / Rupture<br>Block / Cancel / Close / Closure / Impassible / Uninhabitable<br>Buried / Erode / Erosion / Sinkhole<br>Covered with water / Flood / Floodwater / Flow into / Inundate / Overflow / Submerge /<br>Swept away / Washed out • away<br>Crumple / Cut off / Topple / Take the brunt of<br>Damage / Demolish / Destroy / Destruct / Devastate /Disrupt<br>Explosion / Fire<br>Fall across / Knock over • down<br>Filled with sand / Leakage /<br>Malfunction / Outage / Shut down |

**Table 4**
Examples of SQuAD v1.1, NHC, and earthquake datasets.

| Question | Paragraph | Ground truth |
|---|---|---|
| What are the ties that best described what the "eight counties" are based on? (SQuAD v1.1) | Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. … | Demographics and economic |
| What infrastructure in Darlington County has been damaged by Florence and what kind of damage has happened to the infrastructure? (NHC dataset) | … Flooding on the Great Pee Dee River shut down a portion of the city of Florence's municipal water system on September 24. In Darlington County, 23 county maintained roads were damaged due to the hurricane and a bridge on New Hopewell Road collapsed. Flooding damaged approximately 400 homes throughout Dillon County. | In Darlington County, 23 county maintained roads were damaged due to the hurricane and a bridge on New Hopewell Road collapsed |
| What infrastructure in Sukagawa has been damaged by Tohoku earthquake and what kind of damage has happened to the infrastructure? (earthquake dataset) | The Fujinuma irrigation dam in Sukagawa ruptured,[257] causing flooding and the washing away of five homes.[258] Eight people were missing and four bodies were discovered by the morning.[259] [260][261] Reportedly, some locals had attempted to repair leaks in the dam before it completely failed.[262] On 12 March 252 dams were inspected and it was discovered that six embankment dams had shallow cracks on their crests. … | The Fujinuma irrigation dam in Sukagawa ruptured,[257] causing flooding and the washing away of five homes |

running Windows 10 and equipped with an Intel Core i7–7700 and GTX 1080ti. SBERT and BERT models were implemented based on the works of Reimers and Gurevych [43] and Wolf et al. [51], respectively, using Pytorch, an open-source machine learning library developed by Facebook's AI Research Lab.

We used the Stanford Question Answering Dataset 1.1 (SQuAD v1.1) [52], NHC dataset, and earthquake dataset for the experiments. SQuAD v1.1, is a QA dataset produced by the NLP Group of Stanford University and contains 107,785 question–answer pairs. The NHC dataset was collected from 143 National Hurricane Center reports. The earthquake dataset was collected from three Wikipedia articles and five earthquake reports. All questions and answers were generated by the authors, and all paragraphs were manually extracted from the reports. The NHC and earthquake datasets consisted of 533 and 43 question–answer pairs, respectively. In addition, the question–answer pairs in the NHC dataset were used to test the paragraph-retrieval performance of the SBERT model pretrained with the MS MARCO dataset. The paragraph containing the answer is considered as the ground truth in the paragraph retrieval test. The datasets in Excel (.xlsx) file formats can be shared upon request. Table 3 summarizes the datasets, including the type of infrastructure and damage description.

BERT is a pretrained model with a large corpus (BooksCorpus and Wikipedia) derived through NSP and MLM approaches. The pretrained BERT model is first fine-tuned using SQuAD v1.1 for its general QA capability, and then a part of the NHC dataset (435 question–answer pairs from 112 reports) is used to acquire QA capability in the context of infrastructure damage. The effects of SQuAD v1.1 and NHC dataset on fine-tuning was experimentally determined through three cases with SQuAD v1.1 only, NHC dataset only, and both SQuAD v1.1 and NHC dataset. The final fine-tuned model was tested using the remaining 98 question–answer pairs from the 31 NHC reports and 43 from the earthquake dataset for distinct training and test datasets. Table 4 lists the examples of the three datasets (SQuAD v1.1, NHC dataset, and earthquake dataset) used in this study.

### 4.2. Evaluation metrics

Paragraphs with content related to infrastructure damage and specific information on infrastructure damage were separately retrieved. During the retrieval of paragraphs, the SBERT model was assessed using the paragraph evaluation method employed by Chen et al. [45]. This measurement indicates the percentage that the top K retrieved paragraphs contain the ground truth, i.e., containing the answer to the question. For the 533 questions from the NHC dataset, the top 1, 3, 5, and 10 paragraphs related to each question were retrieved from each of the 143 NHC reports. The model was evaluated by investigating whether any of the retrieved paragraphs matched the ground truth. During the retrieval of specific information, the F1-score, which is a model evaluation metric used in the SQuAD v1.1 benchmark, was used. The F1-score is the percentage of the extent to which the ground truth and prediction tokens overlap [52] and is calculated using Eqs. (2–4). The F1-scores for the 98 and 43 question–answer pairs from the NHC and earthquake datasets, respectively, were calculated separately to understand any potential distinctive effects in the two datasets.

**Table 5**
Results of retrieved paragraphs with content related to infrastructure damage (percentage of questions in which ground truth appears in top 1, 3, 5, and 10 retrieved paragraphs).

| Model | Top 1 | Top 3 | Top 5 | Top 10 |
|---|---|---|---|---|
| BERT-base | 4.8 | 34.2 | 53.2 | 80.0 |
| BERT-large | 10.3 | 31.8 | 47.0 | 70.0 |
| SBERT | 59.1 | 80.5 | 85.0 | 90.9 |

**Table 6**
Examples of paragraphs retrieved from hurricane Sandy report (Underlined text: answer to the question).

| Question | Retrieved paragraphs (similarity score) |
|---|---|
| What infrastructure in Hoboken has been damaged by Sandy and what kind of damage has happened to the infrastructure? | … Rescue efforts by the National Guard were required to save residents stranded in the town. About half of the city of Hoboken was reportedly flooded, and at least 20,000 of its residents were surrounded by water at the peak of the surge. The community center in Hoboken, its public works garage, three or four fire houses, and more than 1700 homes were flooded, with damage in the town estimated to be well over $100 million. … (Score: 0.6021) |
| What infrastructure in Florida has been damaged by Sandy and what kind of damage has happened to the infrastructure? | … Persistent northerly winds and the slow movement of Sandy caused very large swells along the east-central and southeastern coasts of Florida. These swells caused moderate to major beach erosion from central Florida southward to Miami-Dade County, along with flooded coastal roadways. Wave heights of up to 20 ft. likely occurred over the Gulf Stream and near shore waters. Wave action caused damage to a stretch of Highway A1A in a portion of the Fort Lauderdale Beach area, and one lane is still closed at the time of this writing. In addition, piers, boat ramps and several coastal homes were damaged from a combination of waves and the high water levels. … (Score: 0.5829) |
| What infrastructure in New Jersey has been damaged by Sandy and what kind of damage has happened to the infrastructure? | … In fact, the extent of catastrophic damage along the New Jersey coast was unprecedented in the state's history, with the brunt of it occurring in Monmouth and Ocean Counties. Whole communities were inundated by water and sand, houses were washed from their foundations, boardwalks were dismantled or destroyed, cars were tossed about, and boats were pushed well inland from the coast. About 5 million residences lost electrical power across this region, with power outages commonly lasting for several weeks. The New Jersey Governor's office estimates that 346,000 housing units were damaged or destroyed in that state, with 22,000 of those units uninhabitable. … (Score: 0.6911) |
| What infrastructure in Manhattan has been damaged by Sandy and what kind of damage has happened to the infrastructure? | … The South Ferry-Whitehall Street station at the southern end of Manhattan was essentially destroyed and subway service between Manhattan and Brooklyn was unavailable for several weeks after the storm. The MTA declared that the overall damage caused by the storm created the worst disaster in the 108-year history of the subway system (e.g. Fig. 30b, 30d). The remainder of New York's transportation infrastructure suffered an estimated $2.5 billion of additional damage. … (Score: 0.5896) |

**Table 7**
Examples of incorrectly retrieved paragraphs.

| Question | Retrieved paragraph (similarity score) | Ground truth (similarity score) |
|---|---|---|
| What infrastructure in Costa Rica has been damaged by Otto and what kind of damage has happened to the infrastructure? | The National Meteorological Institute of Costa Rica reported that ten people perished in flooding in that country as result of Otto. Most of the deaths occurred from flash flooding and landslides over the northern and northwestern portions of Costa Rica. (Score: 0.6123)<br>There have been no official monetary damage estimates received from the affected countries as of this writing. The meteorological service of Costa Rica reported that economic losses are still being calculated, but the report indicated that most of the damage occurred within the agricultural, forestry, and livestock industries. … (Score: 0.5723)<br>… Although the center of the tropical cyclone remained well north of Panama, heavy rainfall and gusty winds associated with Otto appeared to be directly responsible for the fatalities. Three people died in landslides, a young boy perished when a tree fell on his mother's truck, and two youths were swept to their deaths in a swollen river in the eastern portion of Panama City. … (Score: 0.5588)<br>… Otto became a rare Atlantic-to-Pacific basin-crossing tropical cyclone when it moved across southern Nicaragua and northern Costa Rica and emerged over the far eastern North Pacific as a tropical storm. Heavy rainfall and flooding from the hurricane caused 18 fatalities in Central America. (Score: 0.5329)<br>… Otto weakened to a tropical storm just before it exited the Pacific coast of northwestern Costa Rica near the Gulf of Papagayo around 0330 UTC 25 November. … This caused the tropical storm to gradually weaken while it moved west-southwestward at an increasingly faster forward speed. … (Score: 0.5243) | Media reports indicate that extensive freshwater flood damage occurred in portions of Panama, Costa Rica, and Nicaragua, with damage more widespread across northern Costa Rica and extreme southern Nicaragua. Red Cross officials reported that numerous homes, roads, and bridges were damaged in Costa Rica (Fig. 9). Extensive power outages occurred in that country and tens of thousands were left without fresh water after the storm due power outages and damage to numerous water systems. In Panama, Otto had a large effect on the agricultural and livestock industry and media reports indicate that about 500 people were displaced by the storm. More than 10,000 people in Nicaragua were evacuated prior to the hurricane. (Score: 0.4076) |

$$F1 = \frac{2*precision*recall}{precision + recall} \tag{2}$$

$$Precision = \frac{True\ positive\ (TP)}{True\ positive\ (TP) + False\ positive\ (FP)} \tag{3}$$

$$Recall = \frac{True\ positive\ (TP)}{True\ positive\ (TP) + False\ negative\ (FN)} \tag{4}$$

where TP is the number of tokens that overlap between the ground truth and prediction; FP is the number of tokens in the prediction but not in the ground truth, and FN is the number of tokens in the ground truth but not in the prediction.

### 4.3. Results and discussions

Table 5 summarizes the results of retrieving top 1, 3, 5, and 10 paragraphs. The best performance was obtained when using SBERT. In particular, when K was set to 1, the performance of BERT-base and BERT-large was only 4.8% and 10.3%, respectively, whereas SBERT achieved 59.1%. SBERT did not learn the context of infrastructure damage; nevertheless, in retrieving the top 10 paragraphs, relevant paragraphs are retrieved for 90.9% of the questions. Table 6 summarizes some examples in which the top 1 paragraph contains the correct information on infrastructure damage. The underlined portions indicate the infrastructure damage information associated with the question. Table 7 summarizes cases where the retrieved paragraphs did not include the ground truth. For example, the ground truth is not found in the top five paragraphs retrieved on the infrastructure damage caused by Hurricane Otto in Costa Rica. Costa Rica was among the main areas affected, and many descriptions related to Costa Rica were included in the report. The four retrieved paragraphs show high cosine similarity scores because they accounted for the human damage and economic losses incurred in Costa Rica. The third paragraph had a high score, but it pertained to damage in Panama and not in Costa Rica and could have been retrieved because it described the overall damage caused by Hurricane Otto. In contrast, the ground truth listed in Table 7 contains information on the infrastructure damage in Costa Rica (underlined text); however, it has a lower similarity score than the top five paragraphs because the proportion of damage information described is relatively small.

Prior to the experiments on the retrieval of specific information, hyperparameters were determined to fine-tune the BERT model for QA. When fine-tuning BERT, the values of the three hyperparameters (batch size, learning rate, and number of epochs) can be adjusted, and they considerably impact the performance because fewer datasets are used for fine-tuning [4]. The batch size could not be set higher than 1 due to the computer specifications (Intel Core i7–7700 processor and GTX 1080ti GPU) used in this study; hence, only the other two hyperparameters were adjusted. The values of the hyperparameters were determined using five-fold cross-validation. The learning rate and number of epochs for fine-tuning used in this study were 7e−6 and 6, respectively.

Table 8 summarizes the effects of the datasets used for model

**Table 8**
Retrieved infrastructure damage information by training datasets.

| | SQuAD v1.1 | NHC reports | SQuAD v1.1 + NHC reports |
|---|---|---|---|
| F1-score (%) | 33.6 | 74.7 | 90.5 |

**Table 9**
Retrieved infrastructure damage information by test datasets.

| | NHC dataset | Earthquake dataset |
|---|---|---|
| F1-score (%) | 90.5 | 83.6 |

training on the BERT results. The experimental results show that the model fine-tuned using the combined SQuAD v1.1 and NHC dataset exhibited the highest performance, which is 56.9% and 15.8% higher than that with the SQuAD v1.1 and NHC dataset, respectively. These results demonstrated that the BERT model learns the general QA capability from SQuAD v1.1 and the QA capability in the context of infrastructure damage from the NHC dataset. The earthquake dataset is used to perform further experiments on the retrieval of specific infrastructure damage information. The F1-score for the earthquake dataset is 83.6%, as shown in Table 9, despite being trained on the tropical cyclone dataset (NHC reports). Table 10 summarizes the examples in which the F1-score is 100% among the retrieval results with the proposed model.

Among the 141 question–answer pairs used for testing (98 and 43 pairs of the NHC and earthquake datasets, respectively), 115 pairs had an F1-score of 70% or higher. Examples of question–answer pairs with F1-scores of less than 70% are summarized in Table 11. The first example listed in Table 11 depicts the predicament when a paragraph contains two or more answers to a question. The paragraph had two pieces of information regarding the infrastructure damage in Loreto; however, the proposed model only extracted the second piece of information as the answer because of the word "damaged," which is more directly related to the question. In the second example on infrastructure damage in Costa Rica, the model only extracted portions of the answer because it failed to associate Costa Rica with Guanacaste and Punta Arenas. The phrase "in these areas," in the sentence preceding the ground truth, denotes "the cities of Guanacaste and Punta Arenas," which are cities in Costa Rica. The model interpreted the phrase "in these areas" as separate cities not belonging to Costa Rica. For validation, the question was modified to include other location names (e.g., "What infrastructure in Guanacaste has been damaged by Alma and what kind of damage has happened to the infrastructure?"). As expected, the model extracted the following answer: "the hardest hit areas in Costa Rica were the cities of Guanacaste and Punta Arenas. In these areas more than a thousand homes were damaged and 150 were destroyed." The third example shows a case where the model failed to retrieve the correct sentence, although clear words (e.g., "collapsed" and "damaged") were used to denote infrastructure damage. The error could be due to a combination of the following factors. First, the words "collapsed" and "damaged" in the example sentences are not used as nouns, verbs, or participles but as categories to distinguish the degree of damage. Quotation marks to denote categories also had an effect because the usage differed from that in the training dataset. In other words, the model could not extract parts of the answer because the grammar rules differed in the training and test datasets. The second factor is related to the clause, "a report by the National Police Agency of Japan on 10 September 2018 listed." Fig. 4 shows the start scores of tokens with and without the clause; "121" and "the" are the start tokens for the ground truth and prediction, respectively. As shown in Fig. 4, the start score of the token "121" increases significantly when the clause does not exist because no confusing context is represented in the token embedding. The last example shows that only the last three sentences that constitute the ground truth for infrastructure damage in Santiago are retrieved. This is because the model did not recognize the second sentence, which contained information regarding infrastructure damage related to "fire" in a chemical plant as information related to infrastructure damage. Unlike earthquakes, which cause fires or explosions as secondary damage, tropical cyclones generally do not lead to such damage. The fire in a chemical plant resulting from the earthquake was not recognized as damage because we trained the proposed model on the NHC dataset. Consequently, only the third sentence was retrieved because the proposed model can only extract one answer. Furthermore, the damage description in the third sentence (e.g., "have been damaged" and "closed off") was more directly related to the question than that in the first sentence (e.g., "collapsed").

This study aimed to retrieve information on infrastructure damage from disaster-related textual data and provide this information to users,

**Table 10**
Examples of retrieved infrastructure damage information (Underlined text: ground truth; Bold text: prediction).

| Question | Ground truth and prediction in paragraph |
|---|---|
| What infrastructure in Mexico Beach has been damaged by Michael and what kind of damage has happened to the infrastructure? | … **In Mexico Beach, 1584 buildings out of 1692 in the town were reported damaged, with 809 of those reported destroyed**. While exact numbers are not available from the Tyndall AFB, every building was reported damaged with many destroyed. The winds and surge also caused less severe, but extensive, damage elsewhere in the eastern portion of the Panama City metropolitan area. This resulted in a total of more than 45,000 structures damaged and more than 1500 destroyed in Bay County, including severe damage to two hospitals. … |
| What infrastructure in Guatemala City has been damaged by Agatha and what kind of damage has happened to the infrastructure? | The flooding and mudslides also caused an estimated $1.1 billion (US) in property damage, with $982 million in Guatemala and $112 million in El Salvador. A spectacular example of damage documented by the news media was **a 20-m-wide sinkhole that opened up in Guatemala City, destroying several buildings in the process**. |
| What infrastructure in Fukushima Daiichi has been damaged by Tohoku earthquake and what kind of damage has happened to the infrastructure? | … The backup cooling process is powered by emergency diesel generators at the plants and at Rokkasho nuclear reprocessing plant.[296] **At Fukushima Daiichi and Daini, tsunami waves overtopped seawalls and destroyed diesel backup power systems, leading to severe problems at Fukushima Daiichi, including three large explosions and radioactive leakage**. Subsequent analysis found that many Japanese nuclear plants, including Fukushima Daiichi, were not adequately protected against tsunamis.[297] Over 200,000 people were evacuated.[298] … The combined impacts of initial earthquake and its aftershocks resulted in death of 8896 people, injured 22,303 people and impacted the lives of 8 million people. A post-disaster need assessment estimated a total value of disaster damages and losses of approximately 7 billion US dollars. **Some 2656 government buildings and 19,000 classrooms were completely destroyed**. Further details on the damage and impact of earthquake and subsequent aftershocks can be found in the Post Disaster Needs Assessment (PDNA) report of the Government of Nepal's National Planning Commission in 2015. |
| What infrastructure has been damaged by Nepal earthquake and what kind of damage has happened to the infrastructure? | |

**Table 11**
Examples of incorrectly retrieved infrastructure damage information (Underlined text: ground truth; Bold text: prediction).

| Question | Ground truth and prediction in paragraph |
|---|---|
| What infrastructure in Loreto has been damaged by Ivo and what kind of damage has happened to the infrastructure? | In the southern Baja California peninsula, heavy rains associated with Ivo caused Arroyo San Telmo (San Telmo Creek) to overflow its banks, which resulted in the flooding of 200 homes in Loreto. A total of 400 residents were evacuated with the anticipation that the rising water would affect them. Six people were reported injured. **Several roadways near Loreto were damaged**, and six people were injured in a car accident related to the weather in this region. The water supply to the city of Loreto was also cut off temporarily. There is no monetary estimate of damage in the areas affected. |
| What infrastructure in Costa Rica has been damaged by Alma and what kind of damage has happened to the infrastructure? | According to a report from the United Nations (UN) Office for the Coordination of Humanitarian Affairs (OCHA), the hardest hit areas in Costa Rica were the cities of Guanacaste and Punta Arenas. In these areas more than a thousand homes were damaged and 150 were destroyed. **Over 100 roads and bridges in Costa Rica were damaged, which left many communities isolated for several days**. According to some reports, the flooding from Alma in Costa Rica was worse than the flooding experienced from Hurricane Cesar (1996) or Mitch (1998). Monetary losses in Costa Rica are estimated at $33 million U.S. dollars. |
| What infrastructure in Japan has been damaged by Tohoku earthquake and what kind of damage has happened to the infrastructure? | A report by the National Police Agency of Japan on 10 September 2018 listed 121,778 buildings as "total collapsed", with a further 280,926 buildings "half collapsed", and another 699,180 buildings "partially damaged". [51] **The earthquake and tsunami also caused extensive and severe structural damage in north-eastern Japan, including heavy damage to roads and railways as well as fires in many areas, and a dam collapse**. [37,52] Japanese Prime Minister Naoto Kan said, "In the 65 years after the end of World War II, this is the toughest and the most difficult crisis for Japan." [53] Around 4.4 million households in northeastern Japan were left without electricity and 1.5 million without water.[54] |
| What infrastructure in Santiago has been damaged by Chile earthquake and what kind of damage has happened to the infrastructure? | According to an Associated Press Television News cameraman, some buildings collapsed in Santiago and there were power outages in parts of the city. [69] A fire was reported in a chemical plant on the outskirts of Santiago and caused the evacuation of the neighborhood. [28] **Santiago's International Airport seemed to have been damaged and the airport authority closed off all flight operations** for 24 h from around 12:00 UTC. [6] On Sunday, 28 February, Ricardo Ortega, head of the Chilean Air Force, said commercial airline services had been partially re-established and aircraft were being allowed to land in Santiago.[70] |

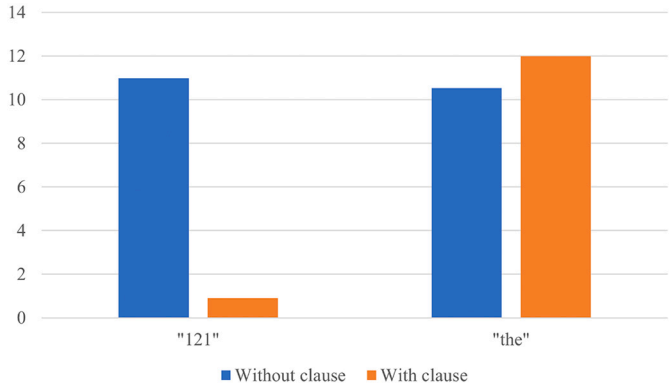


**Fig. 4.** Start scores of the tokens ("121" and "the") with and without the clause.

such as government agencies. Table 9 demonstrates the potential of the proposed method for automating the retrieval of infrastructure damage information from textual data. However, this method has certain limitations, as shown in Tables 7 and 11. Three approaches can be employed as potential solutions. The first is to use a different NLP technique, such as named entity recognition (NER) embedding, with the proposed method. Chowdhury et al. [34] showed that the model performance could be improved with additional information in embeddings. Hence, new information can be added to the words in Table 5 using NER that maps predefined classes to specific words in the text. Using NER information, the language model of the proposed method can focus more on infrastructure damage information to solve the problems listed in Tables 7 and 11. The situation in Table 7 was caused by the little information about infrastructure damage compared to the length of the paragraph. Because NER information increases the weight of infrastructure damage information, the paragraph will achieve a higher similarity score than before. In addition, NER embedding can resolve the problem in the second example in Table 11. The words "collapsed" and "damaged" are classified into types of infrastructure and damage description categories by NER embedding so that the BERT model can accurately extract the answer. The second approach involves training the proposed BERT model with a larger amount of textual data, in addition to the NHC dataset used in this study, to improve the domain knowledge of infrastructure damage. In addition, the fourth example listed in Table 11 shows the necessity of learning about different types of disasters. The last approach involves developing a model with multiple-answer capabilities to resolve problems similar to the first example in Table 11. However, QA with multiple answers remains a challenging task [53].

## 5. Conclusion

In this study, a new method for retrieving infrastructure damage information, such as the type of damaged infrastructure and the type and extent of damage in the affected area from disaster-related textual data is proposed. The method comprises two steps: paragraph retrieval and retrieval of specific information. In retrieving the top 10 paragraphs, the SBERT model pretrained with the MS MARCO dataset retrieved the correct paragraph for 90.9% of the questions. The NHC dataset comprising 435 question–answer pairs was generated from 112 NHC reports to fine-tune the BERT model pretrained with SQuAD v1.1 to retrieve specific information on infrastructure damage. The fine-tuned model was tested with 98 question–answer pairs from the additional 31 NHC reports and 43 question–answer pairs from earthquake-related Wikipedia articles and reports. The fine-tuned model exhibited F1-scores of 90.5% and 83.6% for the tropical cyclone and earthquake data, respectively.

The contribution of this study to infrastructure and disaster management is two-fold. First, we introduce a novel QA method for

automatically retrieving infrastructure damage information from textual data using the BERT model. The QA system can provide detailed information through direct communication between humans and computers. The BERT model can construct accurate QA systems because of its ability to understand the text based on the context. The general BERT model pretrained with Wikipedia and BooksCorpus lacks knowledge regarding infrastructure damage, making the retrieval of precise infrastructure damage information difficult. However, by incorporating domain knowledge, the model can accurately identify the answers to questions related to areas damaged by tropical cyclones and earthquakes. Second, the generated disaster-related dataset for QA comprising 533 question–answer pairs from 148 disaster reports and three Wikipedia articles can be used by future researchers who intend to develop new QA systems for disaster management. The proposed method is expected to aid infrastructure managers in their proactive search for information regarding infrastructure damage sustained in the past while laying plans to reduce infrastructure damage and minimize economic and social losses.

The study successfully demonstrated the potential of the proposed method in infrastructure management related to natural disasters. However, further studies are needed to address these issues. First, more infrastructure and disaster-related textual data should be obtained to develop a robust model. In addition to the hurricanes and earthquakes covered in this study, other natural disasters and extreme weather events that damage infrastructure should also be considered. The acquisition of additional textual data can promote robust models for various disasters and extreme weather events. Second, the top 1 paragraph retrieval should be improved to develop an end-to-end system. The overall accuracy of the proposed method is 53.5%, which is attributed to the performance of the top 1 paragraph retrieval. There is room for significant improvement because the top 3, 5, and 10 paragraph retrieval accuracies are 80.5%, 85.0%, and 90.9%, respectively. Fine-tuning a language model with infrastructure and disaster-related textual data in the paragraph-retrieval step can improve accuracy. Third, multilingual models can be developed for non-English users. The proposed method was verified only for textual data written in English. Multilingual BERT models are available for top 100 popular languages on Wikipedia. It is expected that the proposed method can be applied to other languages through fine-tuning of the models using appropriate datasets. With these improvements, the proposed method can reduce infrastructure management damage caused by natural disasters and extreme weather events.

## Data availability

All the train and test datasets used in this study are available from the corresponding author upon reasonable request.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. 2018R1A6A1A08025348) and the National R&D Project for Smart Construction Technology funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation (No. 21SMIP-A158708-02).

## References

[1] Y. Goldberg, Neural network methods for natural language processing, in: Synthesis Lectures on Human Language Technologies 10(1), 2017, pp. 1–309, https://doi.org/10.2200/S00762ED1V01Y201703HLT037.
[2] Z. Abbasiyantaeb, S. Momtazi, Text-based question answering from information retrieval and deep neural network perspectives: a survey, Wiley Interdiscip. Rev. (2021), https://doi.org/10.1002/widm.1412.
[3] O. Kolomiyets, M.-F. Moens, A survey on question answering technology from an information retrieval perspective, Inf. Sci. 181 (24, 2011) 5412–5434, https://doi.org/10.1016/j.ins.2011.07.047.
[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv Preprint, 2018. arXiv:1810.04805. https://arxiv.org/abs/1810.04805.
[5] NHC (National Hurricane Center), Tropical Cyclone Reports. https://www.nhc.noaa.gov/data/tcr/>, 2020 (Accessed: January 22, 2021).
[6] World Bank, Haiti Earthquake PDNA: Assessment of Damage, Losses, General and Sectoral Needs. https://documents.worldbank.org/curated/en/355571468251125062/Haiti-earthquake-PDNA-Post-Disaster-Needs-Assessment-assessment-of-damage-losses-general-and-sectoral-needs>, 2010 (Accessed: August 20, 2021).
[7] World Bank, Chile - The Magnitude 8.8 Offshore Maule Region Chile Earthquake of February 27, 2010: Preliminary Summary of Damage and Engineering Recommendations. https://documents.worldbank.org/curated/en/750511468217448787/Chile-The-magnitude-8-8-offshore-Maule-region-Chile-earthquake-of-February-27-2010-preliminary-summary-of-damage-and-enginering-recommendations>, 2010 (Accessed: August 20, 2021).
[8] IFRC (International Federation of Red Cross and Red Crescent Societies), Emergency Appeal Final Report Nepal: Earthquake 2015. https://www.ifrc.org/en/publications-and-reports/appeals/?ac=MDRNP008&at=0&c=&co=&dt=1&f=&re=&t=&ti=&zo=>, 2019 (Accessed: January 22, 2021).
[9] World Vision, Nepal Earthquake Response - One Year and Beyond. https://www.wvi.org/nepal/publication/nepal-earthquake-response-one-year-and-beyond>, 2016 (Accessed: January 22, 2021).
[10] RMS (Risk Management Solutions), Y. Xie, Managing Tsunami Risk in the Aftermath of the 2004 Indian Ocean Earthquake & Tsunami. https://www.disastersrus.org/emtools/tsunami/indianoceantsunamireport.pdf>, 2006 (Accessed: August 20, 2021).
[11] M. Al Qady, A. Kandil, Concept relation extraction from construction documents using natural language processing, J. Constr. Eng. Manag. 136 (3) (2010) 294–302, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000131.
[12] J. Lee, J.-S. Yi, J. Son, Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based NLP, J. Comput. Civ. Eng. 33 (3) (2019), 04019003, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000807.
[13] F.U. Hassan, T. Le, Automated requirements identification from construction contract documents using natural language processing, J. Legal Affairs Dispute Resol. Eng. Construct. 12 (2) (2020), 04520009, https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379.
[14] J. Lee, Y. Ham, J.-S. Yi, J. Son, Effective risk positioning through automated identification of missing contract conditions from the contractor's perspective based on FIDIC contract cases, J. Manag. Eng. 36 (3) (2020), 05020003, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000757.
[15] T. Kim, S. Chi, Accident case retrieval and analyses: using natural language processing in the construction industry, J. Constr. Eng. Manag. 145 (3) (2019), 04019004, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001625.
[16] F. Zhang, H. Fleyeh, X. Wang, M. Lu, Construction site accident analysis using text mining and natural language processing techniques, Autom. Constr. 99 (2019) 238–248, https://doi.org/10.1016/j.autcon.2018.12.016.
[17] H. Baker, M.R. Hallowell, A.J.P. Tixier, Automatically learning construction injury precursors from text, Autom. Constr. 118 (2020) 103145, https://doi.org/10.1016/j.autcon.2020.103145.
[18] W. Fang, H. Luo, S. Xu, P.E. Love, Z. Lu, C. Ye, Automated text classification of near-misses from safety reports: an improved deep learning approach, Adv. Eng. Inform. 44 (2020), 101060, https://doi.org/10.1016/j.aei.2020.101060.
[19] D.A. Salama, N.M. El-Gohary, Automated compliance checking of construction operation plans using a deontology for the construction domain, J. Comput. Civ. Eng. 27 (6) (2013) 681–698, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000298.
[20] D.M. Salama, N.M. El-Gohary, Semantic text classification for supporting automated compliance checking in construction, J. Comput. Civ. Eng. 30 (1) (2016), 04014106, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000301.
[21] J. Zhang, N.M. El-Gohary, Automated information transformation for automated regulatory compliance checking in construction, J. Comput. Civ. Eng. 29 (4) (2015), B4015001, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000427.
[22] J. Zhang, N.M. El-Gohary, Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking, J. Comput. Civ. Eng. 30 (2) (2016), 04015014, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346.
[23] P. Zhou, N. El-Gohary, Domain-specific hierarchical text classification for supporting automated environmental compliance checking, J. Comput. Civ. Eng. 30 (4) (2016), 04015057, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000513.
[24] J. Zhang, N.M. El-Gohary, Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking, Autom. Constr. 73 (2017) 45–57, https://doi.org/10.1016/j.autcon.2016.08.027.

[25] X. Xue, J. Zhang, Building codes part-of-speech tagging performance improvement by error-driven transformational rules, J. Comput. Civ. Eng. 34 (5) (2020), 04020035, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000917.
[26] Y. Mo, D. Zhao, J. Du, M. Syal, A. Aziz, H. Li, Automated staff assignment for building maintenance using natural language processing, Autom. Constr. 113 (2020) 103150, https://doi.org/10.1016/j.autcon.2020.103150.
[27] H. Fan, H. Li, Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques, Autom. Constr. 34 (2013) 85–91, https://doi.org/10.1016/j.autcon.2012.10.014.
[28] J.R. Ragini, P.M.R. Anand, V. Bhaskar, Big data analytics for disaster response and recovery through sentiment analysis, Int. J. Inf. Manag. 42 (2018) 13–24, https://doi.org/10.1016/j.ijinfomgt.2018.05.004.
[29] Y. Wang, J.E. Taylor, Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake, Nat. Hazards 92 (2) (2018) 907–925, https://doi.org/10.1007/s11069-018-3231-1.
[30] M. Yu, Q. Huang, H. Qin, C. Scheele, C. Yang, Deep learning for real-time social media text classification for situation awareness – using Hurricanes Sandy, Harvey, and Irma as case studies, Int. J. Digit. l Earth 12 (11) (2019) 1230–1247, https://doi.org/10.1080/17538947.2019.1574316.
[31] H. Hao, Y. Wang, Leveraging multimodal social media data for rapid disaster damage assessment, Int. J. Disaster Risk Reduct. 51 (2020) 101760, https://doi.org/10.1016/j.ijdrr.2020.101760.
[32] Y. Wang, J.E. Taylor, DUET: data-driven approach based on latent dirichlet allocation topic modeling, J. Comput. Civ. Eng. 33 (3) (2019), 04019023, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000819.
[33] S. Kundu, P. Srijith, M.S. Desarkar, Classification of short-texts generated during disasters: a deep neural network based approach, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 790–793, https://doi.org/10.1109/ASONAM.2018.8508695.
[34] J. Ray Chowdhury, C. Caragea, D. Caragea, Keyphrase extraction from disaster-related tweets, in: The World Wide Web Conference, 2019, pp. 1555–1566, https://doi.org/10.1145/3308558.3313696.
[35] W. Wang, K. Stewart, Spatiotemporal and semantic information extraction from Web news reports about natural hazards, Comput. Environ. Urban. Syst. 50 (2015) 30–40, https://doi.org/10.1016/j.compenvurbsys.2014.11.001.
[36] L. Zheng, C. Shen, L. Tang, C. Zeng, T. Li, S. Luis, S.-C. Chen, Data mining meets the needs of disaster information management, IEEE Trans. Hum. Machine Syst. 43 (5) (2013) 451–464, https://doi.org/10.1109/THMS.2013.2281762.
[37] H.-Y. Chan, M.-H. Tsai, Question-answering dialogue system for emergency operations, Int. J. Disaster Risk Reduct. 41 (2019) 101313, https://doi.org/10.1016/j.ijdrr.2019.101313.
[38] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166, https://doi.org/10.1109/72.279181.
[39] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[40] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning Phrase Representations Using RNN Encoder-decoder for Statistical Machine Translation, arXiv Preprint, 2014. arXiv:1406.1078. https://arxiv.org/abs/1406.1078.
[41] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, arXiv Preprint, 2014. arXiv:1409.0473. https://arxiv.org/abs/1409.0473.
[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-%0Aneed.pdf.
[43] N. Reimers, I. Gurevych, Sentence-bert: Sentence Embeddings Using Siamese bert-networks, arXiv preprint, 2019. arXiv:1908.10084. https://arxiv.org/abs/1908.10084.
[44] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: ICML Deep Learning Workshop 2, 2015. https://www.cs.toronto.edu/~gkoch/files/msc-thesis.pdf.
[45] Y. Wo, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural Machine Translation System: Bridging the Gap Between Human and Machine Translation, arXiv preprint, 2016. arXiv:1609.08144. https://arxiv.org/abs/1609.08144.
[46] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter, arXiv Preprint, 2019. arXiv:1910.01108. https://arxiv.org/abs/1910.01108.
[47] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A Human Generated Machine Reading Comprehension Dataset, CoCo@NIPS. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf, 2016.
[48] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading Wikipedia to Answer Open-Domain Questions, arXiv preprint, 2017. arXiv:1704.00051. https://arxiv.org/abs/1704.00051.
[49] S. Wang, M. Yu, X. Guo, Z. Wang, T. Klinger, W. Zhang, S. Chang, G. Tesauro, B. Zhou, J. Jiang, R 3: Reinforced ranker-reader for open-domain question answering, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16712/0.
[50] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.T. Yih, Dense Passage Retrieval for Open-domain Question Answering, arXiv Preprint, 2020.
[51] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, HuggingFace's Transformers: State-of-the-art Natural Language Processing, arXiv preprint, 2019 arXiv:1910.03771, https://arxiv.org/abs/1910.03771.
[52] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ Questions for Machine Comprehension of Text, arXiv Preprint, 2016. arXiv:1606.05250. https://arxiv.org/abs/1606.05250.
[53] M. Zhu, A. Ahuja, D.C. Juan, W. Wei, C.K. Reddy, Question answering with long multiple-span answers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 3840–3849, https://doi.org/10.18653/v1/2020.findings-emnlp.342.