A Question Answering System Applied to Disasters

Xiaojing Guo
Institute of Public Safety Research, Tsinghua University

Xinzhi Wang*
School of Computer Engineering and Science, Shanghai University
wxz2017@shu.edu.cn

Luyao Kou
Institute of Public Safety Research, Tsinghua University

Hui Zhang
Institute of Public Safety Research, Tsinghua University

ABSTRACT

In emergency management, identifying disaster information accurately and promptly out of numerous documents like news articles, announcements, and reports is important for decision makers to accomplish their mission efficiently. This paper studies the application of the question answering system which can automatically locate answers in the documents by natural language processing to improve the efficiency and accuracy of disaster knowledge extraction. Firstly, an improved question answering model was constructed based on the advantages of the existing neural network models. Secondly, the English question answering dataset pertinent to disasters and the Chinese question answering dataset were constructed. Finally, the improved neural network model was trained on the datasets and tested by calculating the F1 and EM scores which indicated that a higher question answering accuracy was achieved. The improved system has a deeper understanding of the semantic information and can be used to construct the disaster knowledge graph.

Keywords

Emergency management, disaster, natural language processing, deep learning.

INTRODUCTION

Disasters are events that impact humans as well as the natural and social environments on which humans depend (Shi et al. 2020). It is of great importance to extract the useful disaster information including the disaster type, the occurring date and time, and the number of injured and killed people from a large number of documents like news articles, announcements, and reports based on which the situation of the disaster can be assessed and the strategies to mitigate the damage of the disaster can be developed (Moel and Alphen 2009). Professionals will write relevant documents when or after a disaster occurs which are very valuable. There are thousands of documents related to a disaster. For example, Wenchuan earthquake is a devastating earthquake that occurred in Sichuan province, China on May 12, 2008 with about 683,000 pieces of data containing "Wenchuan earthquake" on Google[1] and about 67,400,000 on Baidu[2]. However, traditional systems return relevant documents to the question only at a coarse level and users have to devote extra time and energy to searching the entire document for the answer (Hristidis et al. 2010; Cairns et al. 2011). Besides, disasters are a kind of special crisis situation which is characterized by social chaos and disorder, leading to fear, pressure, and shock (Baum 1987). Such incidents are outside the domain of normal daily experience and beyond the direct control of people, impacting people's ability of information processing (Reser 2007; Tombu and Jolicoeur 2003). Considering the limitations of people and traditional systems, we thought about the possibility to apply the question answering system to disasters.

*corresponding author
[1]https://www.google.cn/
[2]https://www.baidu.com/

With the rapid development of computer science, languages can be understood not only by humans but also by computers. The question answering system uses natural language processing techniques to understand questions and documents. In the question answering system, the task of the computer is to find the answer from the passage according to the given question and the passage, similar to the reading comprehension problem in the school exams. Natural language processing is to process, understand, and use human languages like English and Chinese by computers, which is an interdisciplinary subject of computer science and linguistics (Di Felippo and Dias-da-Silva 2009; Collobert et al. 2011). Deep learning provides technical support for most natural language processing tasks. Deep learning builds a common neural network framework that extracts and analyzes the underlying information of the data to enable an automated learning process. The neural network model is a mathematic model that mimics animals' neural networks to conduct calculations on computers. In the field of natural language processing, the neural network model of deep learning can automatically learn good representations or features of the language, thus completing various natural language processing tasks. The question answering system is a kind of neural network model with a specific structure designed for the question answering task. It greatly reduces the environmental interference, namely the possibility that people's thinking is disturbed in disasters which may cause mistakes in searching for disaster information (Reser 2007). Therefore, the question answering system can be well applied to disaster contexts with high efficiency and accuracy (Ojokoh and Adebisi 2019). A knowledge graph can be constructed based on the output of the question answering system. It describes all kinds of entities or concepts and their relationships in the real world. It builds a huge semantic network, with nodes representing entities or concepts and edges representing attributes or relationships, which presents structured information and a more complete picture of knowledge. Therefore, full information and relational facts of entities or concepts can be quickly obtained from the knowledge graph (Pujara et al. 2013; Wang et al. 2014; Lin et al. 2015). It helps to grasp the situation in disasters, based on which the measures to mitigate the impact of disasters can be taken.

The current models of the question answering system trained on the Stanford Question Answering Dataset (SQuAD) are applied to the English text containing a wide range of topics since SQuAD is a reading comprehension dataset built on top of Wikipedia (Rajpurkar, Zhang, et al. 2016; Rajpurkar, Jia, et al. 2018). This paper aims at improving the model by increasing the question answering accuracy and extending the scope of application to domain-specific English question answering and Chinese question answering. Firstly, the characteristics of the existing neural network models were integrated to improve the question answering accuracy of the neural network model. Second, a disaster dataset was constructed for domain-specific question answering by extracting data related to disasters from SQuAD so that the trained model was able to obtain disaster-related information. Third, considering that English and Chinese were both widely-used languages in the world, a Chinese question answering dataset was constructed so that the trained model could be applied to Chinese.

This paper is organized as follows. The model structure of the question answering system is described in Section 2. Then, the training conditions and results are presented in Section 3. Finally, conclusions are drawn in Section 4.

THE QUESTION ANSWERING MODEL IN A DISASTER CONTEXT

With the passage and question as inputs, the question answering system uses natural language processing techniques conducted on computers to output exact answers to questions, eliminating the need to manually find answers from the documents (Stupina et al. 2016). The model structure of the question answering system is introduced in this section. A simple example of the question answering system is shown in Figure 1. The red indexes in Figure 1 are the ordinal values of the words in the passage. With the input of a question and the corresponding passage, the question answering system automatically outputs the answer to the question by calculating the start index and the end index of the answer in the passage. For example, if the passage "On May 12, 2008, Wenchuan county in Sichuan province, was hit by an 8.0 magnitude earthquake. The earthquake began at 2:28:01 p.m. and the epicenter was 19 km deep." and the question "Which county did the earthquake hit?" are put into the system, it will find the exact answer "Wenchuan" from the passage and output the answer. We selected and used the questions and passages containing disaster keywords in the SQuAD dataset whose information was from Wikipedia. The details of the dataset are in Section 4. The answers contain disaster information like the location, time, and severity based on which we can identify and analyze risks in disaster areas, transmit the information in time to inform other people, and put forward proper measures. Figure 2 shows the knowledge graph plotted according to the disaster information in Figure 1.

The basic framework of the question answering neural network model is shown in Figure 3 (Weissenborn et al. 2017). The inputs of the model are the word embeddings of the question and the passage. Question embeddings are n-dimensional vectors corresponding to each word in the question which can represent the features of each word, while passage embeddings are n-dimensional vectors corresponding to each word in the passage. The embeddings used in this paper are detailed in section 3.1.2. The question embeddings and passage embeddings are encoded

Passage: On[1] May[2] 12[3], 2008[4], Wenchuan[5] county[6] in[7] Sichuan[8] province[9], was[10] hit[11] by[12] an[13] 8.0[14] magnitude[15] earthquake[16]. The[17] earthquake[18] began[19] at[20] 2:28:01[21] p.m.[22] and[23] the[24] epicenter[25] was[26] 19[27] km[28] deep[29].		
Question: Which county did the earthquake hit? **Start Index:** 5 **End Index:** 5 **Answer:** Wenchuan	**Question:** Which day did the earthquake happen on? **Start Index:** 2 **End Index:** 4 **Answer:** May 12, 2008	**Question:** Where is Wenchuan county? **Start Index:** 7 **End Index:** 9 **Answer:** in Sichuan province
Question: What was the magnitude of the earthquake? **Start Index:** 14 **End Index:** 14 **Answer:** 8.0	**Question:** When did the earthquake begin? **Start Index:** 20 **End Index:** 22 **Answer:** at 2:28:01 p.m.	**Question:** What was the depth of the epicenter? **Start Index:** 27 **End Index:** 28 **Answer:** 19 km

Figure 1. Examples of the question answering system in a disaster context.

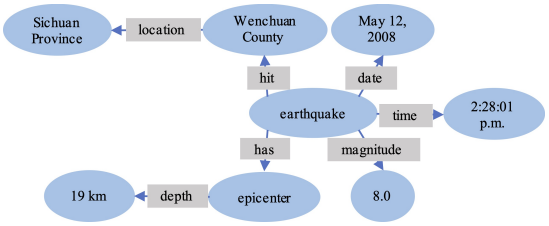

Figure 2. An example of the knowledge graph.

by attention functions to enrich the passage with weighted states from the question and the passage itself. Then the calculation results of the encoder are put into the decoder. Based on the output of the encoder, the decoder calculates the scores of each word in the passage as the start position and the end position of the answer respectively. The index of the word with the highest start position score is the start index of the answer in the passage, and the index of the word with the highest end position score is the end index of the answer in the passage. The words between the start index and the end index (containing the start index and the end index) form the answer to the question. In short, the question and the passage are put into the model, and the encoder and decoder of the model calculate the range of the answer in the passage, outputting the answer text. The encoder and decoder are the core of the model framework which are detailed in section 3.1 and 3.2 respectively.

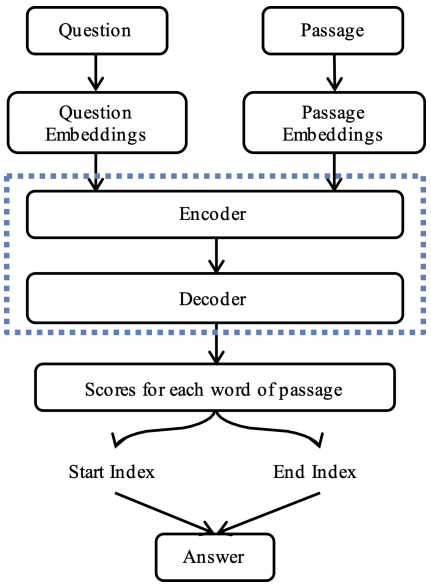

Figure 3. The basic framework of the question answering neural network model.

Based on this basic framework, researchers put forward various innovative models of different advantages. Some of the question answering models and their characteristics are as follows. The r-net model uses a double interaction

layer structure in the encoder. The first interaction layer calculates the interaction between the passage and question by matching the passage and question with gated attention-based recurrent networks, and the second interaction layer calculates the interaction between the words within the passage through a self-matching attention mechanism, which solves the long-term dependency problem of long texts (Group 2017). The Dynamic Coattention Networks (DCN) model is a neural network model for question answering consisting of a coattentive encoder and a dynamic pointer decoder. It uses a dynamic iteration mechanism in the decoder. Each iteration updates the start position and the end position of the answer based on the output of the last iteration. After several rounds of iteration, the answer range becomes more precise and the correct result can be obtained (Xiong et al. 2016). The DCN+ model is developed from the DCN model. It increases the number of attention layers and uses the residual network to fuse the output of each attention layer so that the representation of the passage contains more deep information. At the same time, the self-criticism policy learning mechanism is introduced. It is a concept in reinforcement learning whose reward is related to word overlap between the answer output by the model and the ground truth answer. Therefore, it increases the proportion of the same words between the predicted answer and the ground truth answer (Xiong et al. 2017). The jNet model encodes the question type like when, where, who, how, why, and focuses on the specific information in the passage related to the type of the question, which can better adapt to different types of questions (Zhang et al. 2017). The BERT model (Devlin et al. 2019), which stands for Bidirectional Encoder Representations from Transformers, is a kind of language representation models. It pretrains deep bidirectional representations from unlabeled text based on the Transformer structure (Vaswani et al. 2017) by jointly conditioning on both left and right context in all layers. It uses the masked language model and the next sentence prediction as joint pretraining objectives. It can learn high-quality word representations and be migrated to various natural language processing tasks including question answering. Considering the advantages and disadvantages of the models, this paper combines the encoder layer of the BERT model and the decoder layer of the DCN model to complete the question answering task in a disaster context. The combination performs well and achieves a high question answering accuracy, which is presented in the following sections.

Text Encoding in A Disaster Context

The BERT model is applied for text encoding corresponding to the Encoder in Figure 3. It encodes the question and the passage. The full name of BERT is "Bidirectional Encoder Representations from Transformers". The masked language model and next sentence prediction are used as a joint goal to pre-train the model and get the high-quality representation of words. The model structure, input vectors, and pre-training tasks of BERT are specifically described below (Devlin et al. 2019).

The Model Structure of BERT

The model structure of BERT is shown in Figure 4. The word embeddings E of the question and the passage are put into the model, and the output vector T corresponding to each word can be calculated. The output vector T represents the dependence of the word on the context.

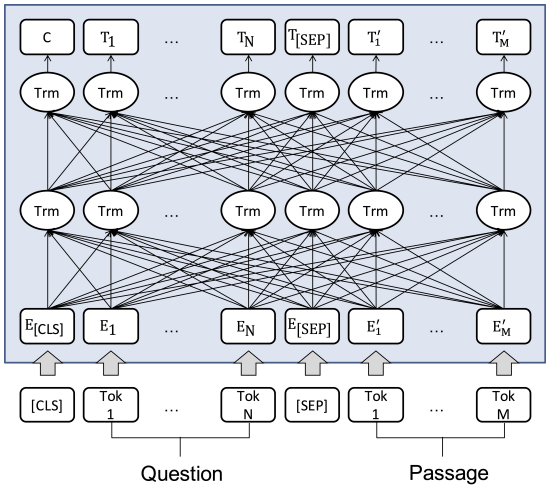

Figure 4. The model structure of BERT.

The Transformer (Trm) is the most important component of the BERT model structure, corresponding to the Trm module in Figure 4. The Transformer structure is the first sequence transduction model based completely on

attention. It converts the distance between any two words in a sentence to 1, effectively solving the long-term dependency problem in natural language processing tasks. A Transformer is composed of several Encoders. Each Encoder consists of a multi-head attention layer and a fully connected layer. It can calculate the interdependence between texts and convert the input text into its corresponding feature vectors. The detailed architecture, equation, and description of the Transformer can be seen in (Vaswani et al. 2017).

The Input Vector of BERT

An example of the input vectors (namely input embeddings) of BERT is shown in Figure 5 where input embeddings are denoted as E. The question and passage in the example are from Figure 1. The input vector of BERT is the sum of three vectors embedded with features, which are as follows.

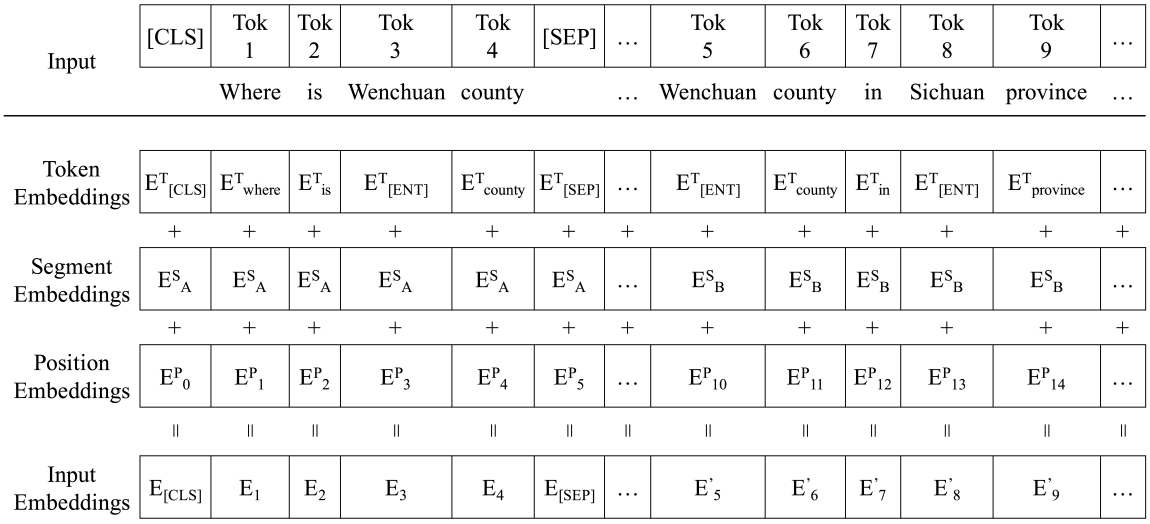

Figure 5. An example of the input vectors of BERT.

Token Embedding: It uses WordPiece (Wu et al. 2016) to characterize the semantics of words. WordPiece can effectively reduce the influence of word suffixes without actual semantics (such as different forms of verbs in different tenses) by dividing words into tokens, namely common character combination units with a high frequency of occurrence. For example, "played", "plays", and "playing" are different tenses of the verb "play" but have the same meaning. WordPiece divides "played" into "play" and "##ed", "plays" into "play" and "##s", and "playing" into "play" and "##ing" and finds the corresponding embeddings E^T_{play}, $E^T_{\#\#ed}$, $E^T_{\#\#s}$, and $E^T_{\#\#ing}$ of the tokens in its token vocabulary. In this way, "played", "plays", and "playing" have the common token "play" and their meanings can be consolidated. Therefore, WordPiece reduces the number of words in common vocabularies and makes the semantics of the words clearer.

Segment Embedding: For the question answering task, the input of the model is not a single sentence, but a question and a passage. A special symbol [SEP] is inserted between the two sentences to separate and distinguish them. Besides, different segment embeddings are assigned to them, that is, the tokens in the question are given the segment embeddings of E^S_A, and the tokens in the passage are given the segment embeddings of E^S_B.

Position Embedding: The change of the token position may completely change the meaning of a sentence, so the position representation is very important for the understanding of a sentence. The position information is encoded into the position embedding to represent the positional relationship between different tokens in the sentence. For example, the first token has the position embedding of E^P_0 and the second token has the position embedding of E^P_1.

The Pre-training Tasks of BERT

BERT obtains the pre-trained parameters of the model through two self-supervised pre-training tasks and learns the feature representation of the language so that other natural language processing tasks can be trained based on the pre-trained model, which greatly reduces the training time and computational load.

The first pre-training task is the Masked Language Model. To obtain the deep two-way representation of the language, some words in the text are randomly masked, and the masked words are predicted by BERT according to the left and right context, similar to a cloze test. 15% of the words from all the training data are randomly selected

as the masking objects. However, considering the specific circumstances of the training process, the selected words are not masked with a 100% probability. The specific practices are described below.

80%: replace the selected word with [MASK].

10%: replace the selected word with a random word.

10%: keep the selected word unchanged.

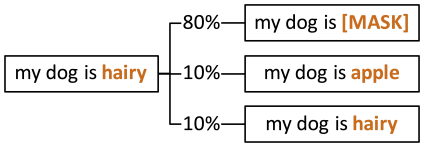

Figure 6. An example of the Masked Language Model.

An example of the Masked Language Model is shown in Figure 6. In this way, the BERT model does not know which word it needs to predict, nor which word is replaced by a random word, so it must maintain the uniform contextual representation for each word. Besides, since only 1.5% of the words are randomly replaced (10% * 15%), the language comprehension ability of the model is not damaged.

The second pre-training task is the next sentence prediction. The question answering task is based on the understanding of the relationship between the question and the passage, but the masked language model cannot learn the relationship between sentences. The next sentence prediction task is designed to train a model that understands the relationship between sentences.

The task of the next sentence prediction is to determine whether sentence B is the next sentence of sentence A. If B is the next sentence of A, the output is "IsNext"; if B is not the next sentence of A, the output is "NotNext". The training data is obtained by randomly extracting two consecutive sentences from the parallel corpus, in which 50% of the data retains the two sentences, and the other 50% replaces the second sentence with a randomly extracted sentence. Next sentence prediction is of great significance for improving the accuracy of the question answering task.

Text Decoding in A Disaster Context

The DCN model is applied for text decoding corresponding to the Decoder in Figure 3. The output vectors T' of the encoder corresponding to each word in the passage are decoded by the Dynamic Pointer Decoder of the Dynamic Coattention Networks (DCN) model (Xiong et al. 2016). The Dynamic Pointer Decoder imitates the process of humans repeatedly thinking about the solution of the practical problems and uses a dynamic iterative mechanism to update the start and end positions of the answer until the positions do not change or the set number of iterations is reached. The model structure of Dynamic Pointer Decoder is shown in Figure 7. During iteration i, s_i denotes the estimated start index and e_i denotes the estimated end index. T'_{s_i} and T'_{e_i} are the corresponding output vectors of the encoder. The elements related to the estimation of the start position are in blue and the ones related to the end position are in red. Each iteration puts the output matrix T' of the encoder, the calculation results s_{i-1} and e_{i-1} of the last iteration, and the history information h_i into the Highway Maxout Network (HMN). The model structure of HMN is shown in Figure 8. HMN calculates the start score α of each word in the passage and selects the word corresponding to the highest start score (which is "in" in Figure 7) as the start position of the answer. Similarly, another HMN with the same structure and different parameters calculates the end score of each word and selects the word with the highest end score (which is "province" in Figure 7) as the end position of the answer.

HMN combines the Highway Network with the Maxout activation function. The Highway Network connects each layer of the network to the last layer to increase the depth of the model. The Maxout activation function is a kind of activation functions in neural network models, the equation of which is

$$f(x) = max(w_1^T x + b_1, w_2^T x + b_2, ..., w_n^T x + b_n). \quad (1)$$

It gets n outputs per node which are $w_1^T x + b_1, w_2^T x + b_2, ..., w_n^T x + b_n$ and takes the maximum of them as the final output $f(x)$, which can theoretically fit any activation function. The combination of the Highway Network and the

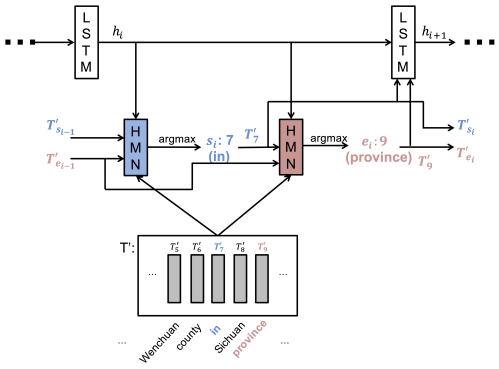

Figure 7. The model structure of Dynamic Pointer Decoder.

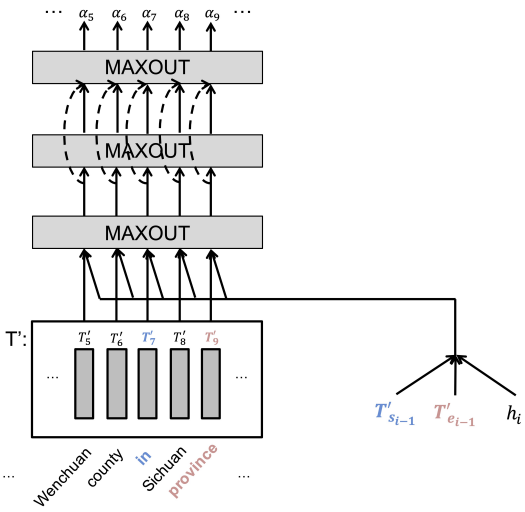

Figure 8. The model structure of Highway Maxout Network.

Maxout activation function improves the performance of the network. The equations and more details of HMN are in the DCN paper (Xiong et al. 2016).

The BERT model and the DCN model have different advantages. The deep bidirectional encoder and the pre-training tasks of the BERT model can learn high-quality word representations that fully fuse contextual information. The dynamic iterative decoder of the DCN model repeatedly updates the calculation results through an iterative process, which helps to get a more accurate answer scope in the passage. Combining the encoder portion of the BERT with the decoder portion of the DCN and training it on a large-scale corpus can improve the accuracy of the question answering task. The overall network structure of the question answering model in a disaster context is shown in Figure 9.

MODEL TRAINING AND RESULT ANALYSIS

The accuracy of the model is highly relevant to the quality of the dataset on which the model is trained. Three datasets are used to train the model. First, the model is trained on the whole dataset of SQuAD 2.0 (Rajpurkar, Jia, et al. 2018), which is a reading comprehension dataset with questions proposed according to Wikipedia articles and answers that are part of the reading passages. Second, a disaster dataset is constructed based on SQuAD 2.0 and all of its content is related to disasters. Third, a Chinese dataset is converted from WebQA (Li et al. 2016), which is a large-scale manual labeled Chinese dataset mainly based on a Chinese question answering website. The training and test of the model on the three datasets are described below.

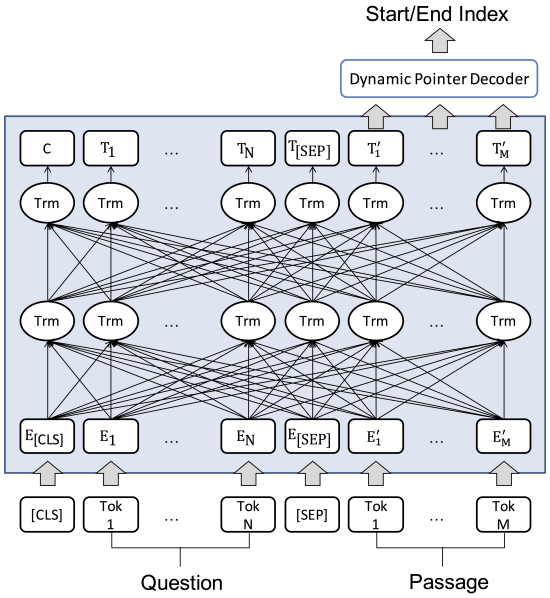

Figure 9. The improved model combining the network structure of BERT and DCN.

Training and Result Analysis on the English Dataset

Training and Result Analysis on the Whole Dataset

The improved model is trained on the question answering dataset and the loss and accuracy information is recorded and evaluated. The Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset built on top of Wikipedia (Rajpurkar, Zhang, et al. 2016). SQuAD consists of passages, questions, and answers. The passage is from the Wikipedia articles, the question is presented by the labeling personnel based on the passage, and the answer is the fragment of the passage. There are two versions of SQuAD currently. SQuAD 1.1 contains more than 100,000 question-answer pairs based on over 500 articles.

SQuAD 2.0 (Rajpurkar, Jia, et al. 2018) adds more than 50,000 questions that do not have answers in passages to SQuAD 1.1. SQuAD 2.0 puts forward higher requirements for the question answering system, that is, not only should it find the answer from the passage according to the question, but also judge whether the question can be answered. SQuAD 2.0 well suits the situation in disasters because it is impossible to collect all the disaster-related information in a short time. Except for obtaining disaster information timely and accurately, the question answering system should also indicate which information is missing or not yet known to provide guidance for information collection and emergency rescue of the next stage. Considering that, the disaster model is firstly trained and tested on the SQuAD 2.0 dataset.

SQuAD 2.0 contains 130,000 pieces of training data and more than 10,000 pieces of test data. The improved neural network model based on BERT and DCN was trained on SQuAD 2.0 using 4 NVIDIA Tesla V100s with a batch size of 12, a learning rate of 5e-5, and a training epoch of 19.

The model calculates a start score and an end score for each word in the passage. In the example shown in Figure 10, the word with the highest start score is "in" whose index is 7, and the word with the highest end score is "province" whose index is 9. Therefore, the predicted answer is "in Sichuan province" which is consistent with the ground truth answer.

Question:	Where is Wenchuan county?						
Answer:	in Sichuan province						
Passage:	…	Wenchuan	county	in	Sichuan	province	…
Index:	…	5	6	7	8	9	…
Start Score:	…	0.812	0.633	7.126	3.754	0.299	…
End Score:	…	0.307	0.550	0.625	2.925	9.341	…
Prediction:				in	Sichuan	province	

Figure 10. An example of scoring.

Loss and accuracy are the indicators of deep learning effects. The loss refers to the error between the predicted result and the ground truth result. The accuracy includes two evaluation indicators for the question answering task which are Exact Match (EM) and F1-score (F1). EM measures the percentage of predictions that exactly match any ground truth answer and F1 measures the average overlap between ground truth answers and predictions. The ground truth answers and predictions are taken as bags of tokens and then their F1 are calculated. The F1 for all the questions in a dataset is computed by taking the maximum F1 for a given question over all of the ground truth answers and then averaging over all the questions. For example, the ground truth answers to the questions in Figure 1 are "Wenchuan", "May 12, 2008", "in Sichuan province", "8.0", "at 2:28:01 p.m.", and "19 km" respectively. If the predicted answers are "Sichuan", "May 12", "in Sichuan province", "8.0", "at 2:28:01 p.m.", and "19 km" respectively, the first predicted answer is totally wrong and two-thirds of the words in the second ground truth answer are the same as the words in the predicted answer. Therefore, for the six questions, EM = $4/6 = 66.7\%$ and F1 = $(0 + 2/3 + 1 + 1 + 1 + 1)/6 = 77.8\%$.

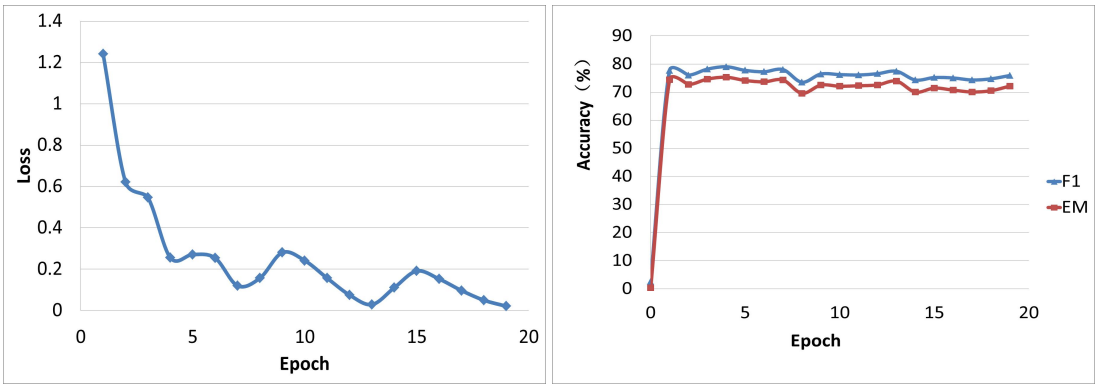

Figure 11. The variation of the loss and accuracy with the epoch on the SQuAD dataset.

In the training scenario of this section, the loss and accuracy (EM, F1) as a function of the training epoch are shown in Figure 11. It can be seen that the accuracy rate reaches the highest at the fourth epoch. The accuracy of the model on the SQuAD 2.0 test dataset can reach F1 = 79.0% and EM = 75.3%. With the same training parameters, the accuracy of the BERT model is F1 = 76.3% and EM = 73.1%. F1 of the improved model is 3.5% higher than the BERT model relatively and EM is 3.0% higher relatively. The DCN model can only be applied to the SQuAD 1.1 dataset and cannot handle the problem in SQuAD 2.0 that the answer cannot be found in the passage. It can be seen that the improved model has a better performance and can well suit the complex situation in a disaster context.

Training and Result Analysis on the Disaster Dataset

We have evaluated the performance of the model on the SQuAD 2.0 dataset and get good results. A disaster training dataset and a disaster test dataset are constructed to further improve the ability of the model to understand the information more closely related to disasters. First, we searched for documents related to disasters in online newspapers and the Internet search engine and collected the words that appeared with high frequency as disaster keywords, considering that high-frequency words had a close relationship with disasters. The keywords are mostly names of all kinds of disasters, like earthquake, flood, plague, drought, hurricane, tornado, typhoon, and so on. There are also words which are probably results of disasters, like die, death, pollute, pollution, and so on. Second, the data containing the keywords of disasters were extracted from the SQuAD 2.0 dataset to form a disaster dataset. The training set of the disaster dataset contains nearly 70,000 pieces of data and the frequencies of the keywords are shown in the left of Figure 12. Among them the five keywords with the highest frequency are ice, war, die, rain, and death and the frequencies are 29055, 23441, 14025, 8072, and 4562; the five keywords with the lowest frequency are tsunami, blizzard, blow up, pestilence, and avalanche and the frequencies are 13, 10, 9, 5, and 3. The test dataset contains nearly 7000 pieces of data and the frequencies of the keywords are shown in the right of Figure 12. Among them the top five high-frequency keywords are ice, war, die, rain, and plague and the frequencies are 2851, 1867, 1601, 881, and 241; the five keywords with the lowest frequency are hail, pollute, tornado, pestilence, and typhoon and the frequencies are 10, 10, 10, 9, and 9 respectively.

The data structure of the disaster training set is shown in Figure 13. The dataset contains several $\{title, paragraphs\}$ pairs. "Title" corresponds to the title of the article. An article can be divided into several pieces to avoid the passage being too long, so each "paragraph" contains several $\{qas, context\}$ pairs. "Context" is the fragment of the article, which is, in other words, the passage corresponding to the question and answer. "Qas" contains several $\{question, id, answers, is\_impossible\}$ quads. "Question" is based on the context and "id" is the unique

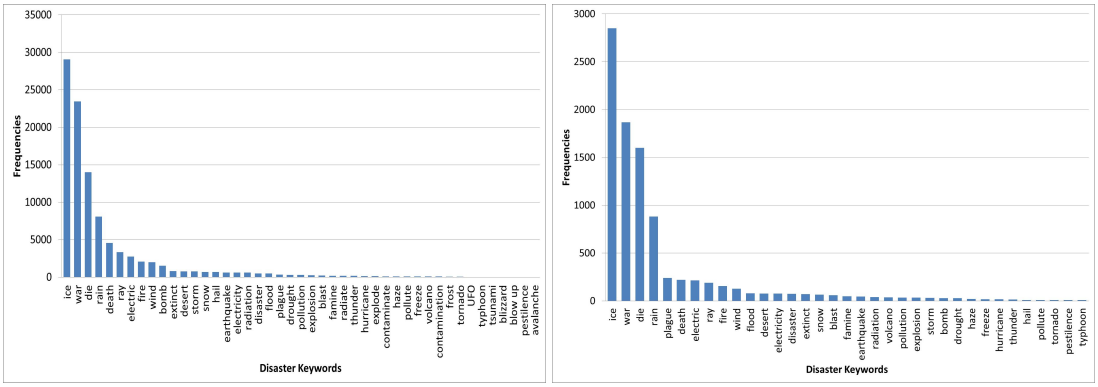

Figure 12. The frequencies of keywords on the disaster training (left) and test (right) dataset.

number of each question. "Is_impossible" is a variable that indicates whether the question can be answered. When the answer can be found from the passage, the value of "is_impossible" is false; when the answer cannot be found from the passage, the value of "is_impossible" is true. If the value of "is_impossible" is false, there is a $\{text, answer\_start\}$ pair in "answers". "Text" is the text content of the answer. "Answer_start" is the start position of the answer in the passage, that is, the position number of the first character of the answer in the passage. If the value of "is_impossible" is true, "answers" will be empty. Instead, an element named "plausible_answers" will be added, which contains a seemingly reasonable, but actually wrong answer. That is the whole frame of the disaster training set. Several examples of the data in the disaster training set are shown in Figure 14. The only difference between the test set and the training set is that there are three $\{text, answer\_start\}$ pairs instead of a $\{text, answer\_start\}$ pair in "answers" of the test set. When calculating the accuracy of question answering on the test set, the predicted answer can be determined to be correct as long as it is the same as any of the three given answers.

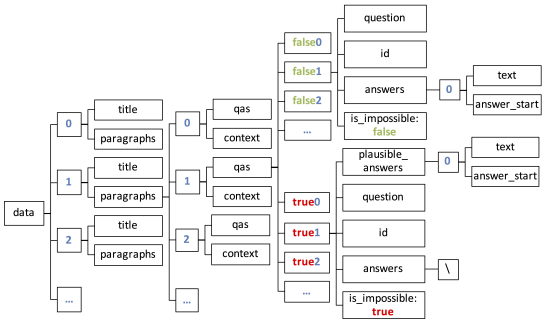

Figure 13. The structure of the disaster training set.

The improved neural network model based on BERT and DCN was trained on the disaster dataset. The GPU used was 4 NVIDIA Tesla V100s, the batch size was 12, the learning rate was 5e-5, and the number of training epochs was 14.

In the training scenario of this section, the loss and accuracy (EM, F1) as a function of the epoch are shown in Figure 15. It can be seen that the accuracy rate reaches the highest at the second training epoch. The accuracy of the model on the disaster test dataset can reach F1 = 78.7% and EM = 75.7%. Compared with the accuracy on the whole dataset (F1 = 79.0%, EM = 75.3%), F1 is 0.4% relatively lower, and EM is 0.5% relatively higher. Although the data volume of the disaster dataset is smaller than the whole dataset, the overall difference in accuracy is little, indicating that the model can capture the linguistic features in the specific field of disaster when trained on the disaster dataset. Therefore, the model is able to extract useful information from disaster-related documents in a disaster context. After the second epoch, the accuracy fluctuated downwardly, indicating that with the increase of the training epochs, although the model can better fit the training data, its transfer ability to the test dataset weakens. More training epochs do not often bring about better results, so the number of epochs should be selected according to the training and testing conditions.

```
{
    "title": "Black_Death",
    "paragraphs": [
        {
            "qas": [
                {
                    "question": "Where did the black death originate?",
                    "id": "5726468470898414009 4c123",
                    "answers": [
                        {
                            "text": "Central Asia",
                            "answer_start": 68
                        }
                    ],
                    "is_impossible": false
                },
                {
                    "question": "How much of the European population did the black death kill?",
                    "id": "5726468470898414009 4c125",
                    "answers": [
                        {
                            "text": "30–60%",
                            "answer_start": 381
                        }
                    ],
                    "is_impossible": false
                },
                {
                    "plausible_answers": [
                        {
                            "text": "1343",
                            "answer_start": 146
                        }
                    ],
                    "question": "In what year did the Black Death originate in Central Asia?",
                    "id": "5a2eb84ba83784001a7d2 41f",
                    "answers": [ ],
                    "is_impossible": true
                }
            ],
            "context": "The Black Death is thought to have originated in the arid plains of Central Asia, where it then travelled along the Silk Road, reaching Crimea by 1343. From there, it was most likely carried by Oriental rat fleas living on the black rats that were regular passengers on merchant ships. Spreading throughout the Mediterranean and Europe, the Black Death is estimated to have killed 30–60% of Europe's total population. In total, the plague reduced the world population from an estimated 450 million down to 350–375 million in the 14th century. The world population as a whole did not recover to pre-plague levels until the 17th century. The plague recurred occasionally in Europe until the 19th century."
        }
    ]
}
```


Figure 14. Examples of the data in the disaster training set.

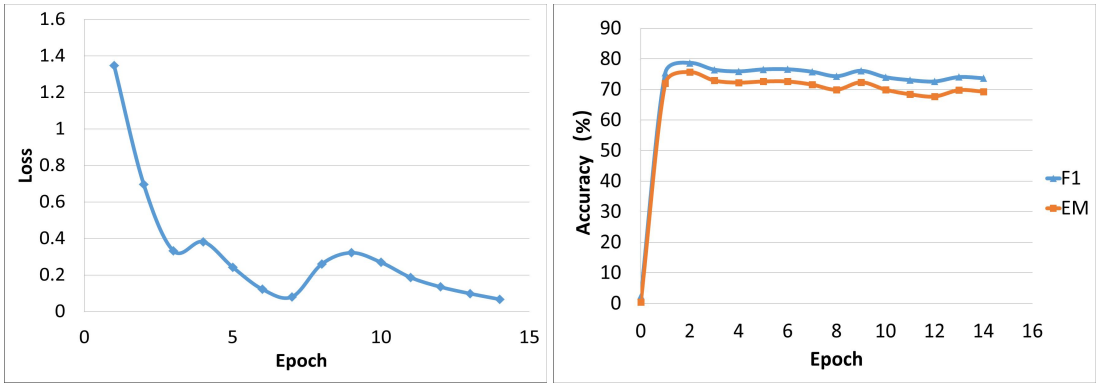

Figure 15. The variation of the loss and accuracy with the epoch on the disaster dataset.

Training and Result Analysis on the Chinese Dataset

Chinese is also a widely-used language besides English. As a country with a large population and broad land and consequently a high incidence of disasters, China is attaching great importance to emergency management (Zou and Yuan 2010; Ye et al. 2012). The above research on the question answering system studies the English language which has different characteristics from the Chinese language (Huang and Yao 2003). If the question answering model can be applied to the Chinese language as well, it will increase the accuracy and timeliness of disaster information extraction in China. So this paper constructs a Chinese question answering dataset based on WebQA (Li et al. 2016).

WebQA is a large-scale manual labeled Chinese dataset based on the Baidu Knows website and other resources which is also composed of $\{passage, question, answer\}$ triples. Most of the data were collected from the large community question answering site, "Baidu Knows", and a small portion was manually collected from the documents of other sites. WebQA also has a large number of data that the answer can be found in the passage. We extracted such data, calculated the start position of the answer in the passage, and converted the data into the same format as the English dataset. Then it could be put into the neural network model for training and testing.

The model should be modified to adapt to the change of the language. The Chinese and English data processing methods are different in terms of the token embedding. In the English text, the words are firstly separated by spaces and then divided by WordPiece into common character combination units. While in the Chinese text, the language is continuous instead of being separated by spaces. One Chinese character is a minimal semantic unit, so the sentence can be split directly into characters. Figure 16 shows the process of turning the Chinese text into input vectors. All common words in Chinese form a vocabulary list and each word has a corresponding serial number in the vocabulary list. The input ids are the serial numbers and the segment ids and position ids are the same as those in the encoding method of BERT.

A training dataset containing more than 30,000 pieces of data and a test dataset containing nearly 3,000 pieces of data were extracted from WebQA. The improved neural network model based on BERT and DCN was trained on

tokens	[CLS]	白	衣	天	使	指	的	是	哪	种	职	业	?	[SEP]
input ids	101	4635	6132	1921	886	2900	4638	3221	1525	4905	5466	689	136	102
segment ids	0	0	0	0	0	0	0	0	0	0	0	0	0	0
position ids	0	1	2	3	4	5	6	7	8	9	10	11	12	13

tokens	白	衣	天	使	即	是	穿	白	大	褂	的	护	士	。	[SEP]
input ids	4635	6132	1921	886	1315	3221	4959	4635	1920	6184	4638	2844	1894	511	102
segment ids	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
position ids	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28

Figure 16. An example of the Chinese input vector.

the Chinese dataset. The GPU used was 4 NVIDIA Tesla V100s, the batch size was 48, the learning rate was 5e-5, and the training epoch was 17.

In the training scenario of this section, the loss and accuracy as a function of the training epoch are shown in Figure 17. It is shown that similar to the above training scenarios, the model reaches high accuracy after one training epoch based on the pre-training parameters, which indicates that pre-training is an effective way to raise the convergence speed as well as reduce computing time and computing load. The accuracy rate reaches the highest at the fourth epoch. The accuracy of the model on the Chinese test dataset can reach 79.9%, which is higher than the accuracy on the English dataset (F1 = 79.0%, EM = 75.3%). Therefore, the model can be well transferred to the Chinese question answering task and can contribute to information extraction in a disaster context in China.

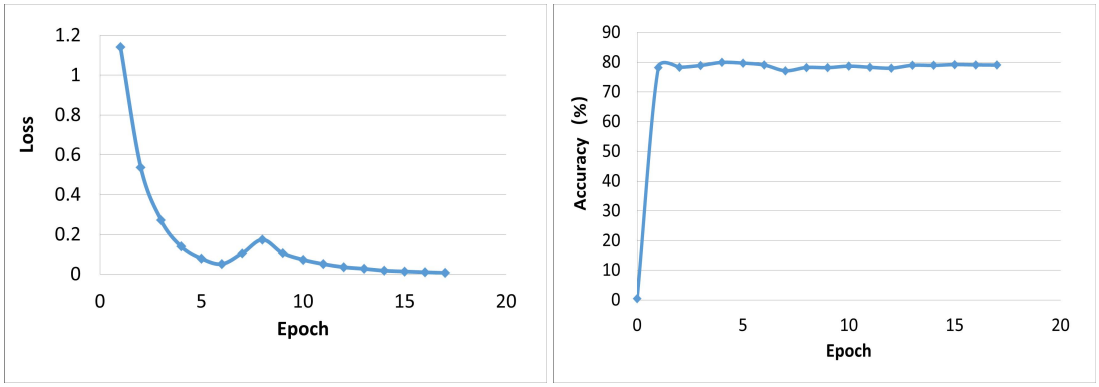

Figure 17. The variation of the loss and accuracy with the epoch on the Chinese dataset.

In summary, the training results of the model on the SQuAD 2.0 dataset, disaster dataset, and Chinese dataset are presented in Table 1.

Table 1. Summary of the results on the three datasets

Dataset	Accuracy
The SQuAD 2.0 dataset	79.0% F1, 75.3% EM
The disaster dataset	78.7% F1, 75.7% EM
The Chinese dataset	79.9%

CONCLUSIONS

The efficiency of the situation assessment and decision making in disasters highly relies on information acquisition, because the decision should be made based on the effective information found from the disaster documents. While traditional systems can return the documents pertinent to the disaster, it takes extra effort for users to read the entire document to locate the disaster information. The employment of computer technology can solve this problem. This paper constructs a question answering system that can be applied to the disaster context to extract information automatically and promote communication. Since the model of the question answering system is trained on datasets, the accuracy of the question answering system depends on the quality of the datasets and the model structure of the system. Therefore, dataset acquisition and model construction are the two main tasks of the question answering system.

In the aspect of dataset acquisition, the contribution of this paper is as follows. First, this paper selects the questions and answers related to disasters and forms the disaster question answering dataset which contains about 70,000 pieces of training data and 7,000 pieces of test data. Second, this paper constructs the Chinese question answering dataset that can be put into the question answering model by modifying the text segmentation method. The Chinese dataset contains more than 30,000 pieces of training data and nearly 3,000 pieces of test data.

In the aspect of model construction, this paper comprehensively analyzes the advantages and disadvantages of the existing models of the question answering system. The BERT model can fully integrate contextual information and calculate the relevance degree of the text. The DCN model can combine the last prediction result and the historical information to iteratively update the answer span. The advantages of BERT and DCN are combined to construct an improved model. F1 and EM are generally accepted metrics to evaluate model performance for question answering (Rajpurkar, Zhang, et al. 2016). All question answering models use the two numbers to describe their accuracies, so the models in our paper use them as well. The improved model is trained and tested on SQuAD 2.0 dataset, the disaster dataset, and the Chinese dataset respectively and the accuracies are calculated. The results are as follows. First, the accuracy achieves F1 = 79.0% and EM = 75.3% on SQuAD 2.0 and is higher than the BERT model which is F1 = 76.3% and EM = 73.1%. The increase in accuracy indicates that our model of the question answering system is able to understand the natural language more deeply. Second, the model of the question answering system trained on the disaster dataset achieves the accuracy of F1 = 78.7% and EM = 75.7%. It can effectively extract disaster-related information from the documents. Third, the accuracy on the Chinese dataset can reach 79.9%, indicating that the model of the question answering system can be transferred to the Chinese disaster context.

In conclusion, we developed a question answering system combining the advantages of existing neural network models. The system was trained on the English and Chinese datasets and tested by calculating the F1 and EM scores which indicated that a high question answering accuracy was achieved. The results show that the question answering system can efficiently extract disaster information such as the location, date, time, and magnitude from the documents related to the disaster consisting of news articles, announcements, reports, and so on based on which a knowledge graph can be constructed. It is practically valuable for learning about the situation and taking measures to mitigate the negative effects in a disaster context. The study has several limitations. First, English and Chinese point to the culture of two types of civilization. Therefore, words can be used in different manners related to these two cultures. Although the similarity of the question answering accuracies of the system on the English and Chinese datasets can in some degree reveal that our approach could adapt to these two cultures, we will go deeply into the different manners of word using related to English and Chinese in the future work and analyze if our approach can show the differences. Second, there are works on pragmatics and interaction analysis domain that study discussions and communication (Youn 2015; Hoque et al. 2018). They may help to extend our approach to an analysis that goes beyond word classification. Nevertheless, the question answering system makes the computer understand implicit semantics instead of simple word matching, which increases the precision rate and recall rate. It can automatically obtain disaster information by computers, saving the energy and time of human beings.

ACKNOWLEDGMENTS

This work is supported by National Key R&D Program of China (No. 2020YFA0714500), National Science Foundation of China (Grant No. 7204100828), Shanghai Sailing Program (No.20YF1413800), High-tech Discipline Construction Fundings for Universities in Beijing (Safety Science and Engineering) and Beijing Key Laboratory of City Integrated Emergency Response Science.

REFERENCES

Baum, A. (1987). "Toxins, technology, and natural disasters." In: *American Psychological Association Convention, Aug, 1986, Washington, DC, US; This chapter is based upon one of the 1986 Master Lectures that were presented at the aforementioned convention.* American Psychological Association.

Cairns, B. L., Nielsen, R. D., Masanz, J. J., Martin, J. H., Palmer, M. S., Ward, W. H., and Savova, G. K. (2011). "The MiPACQ clinical question answering system". In: *AMIA annual symposium proceedings*. Vol. 2011. American Medical Informatics Association, p. 171.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (AUG 2011). "Natural Language Processing (Almost) from Scratch". In: *JOURNAL OF MACHINE LEARNING RESEARCH* 12, 2493–2537.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Di Felippo, A. and Dias-da-Silva, B. C. (SEP-DEC 2009). "Natural language processing as human language engineering". In: *CALIDOSCOPIO* 7.3, 183–191.

Group, N. L. C. (May 2017). "R-NET: Machine Reading Comprehension with Self-matching Networks". In:

Hoque, E., Setlur, V., Tory, M., and Dykeman, I. (JAN 2018). "Applying Pragmatics Principles for Interaction with Visual Analytics". In: *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS* 24.1, 309–318.

Hristidis, V., Chen, S.-C., Li, T., Luis, S., and Deng, Y. (2010). "Survey of data management and analysis in disaster situations". In: *Journal of Systems and Software* 83.10, pp. 1701–1714.

Huang, G. T. and Yao, H. H. (2003). "A System for Chinese Question Answering". In:

Li, P., Li, W., He, Z., Wang, X., Cao, Y., Zhou, J., and Xu, W. (2016). "Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering". In: *CoRR* abs/1607.06275. arXiv: 1607.06275.

Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). "Learning entity and relation embeddings for knowledge graph completion". In: *Twenty-ninth AAAI conference on artificial intelligence*.

Moel, H. and Alphen, J. (Mar. 2009). "Flood maps in Europe–methods, availability and use". In: *Natural Hazards and Earth System Sciences* 9, pp. 289–301.

Ojokoh, B. and Adebisi, E. (Jan. 2019). "A Review of Question Answering Systems". In: *Journal of Web Engineering* 17, pp. 717–758.

Pujara, J., Miao, H., Getoor, L., and Cohen, W. (2013). "Knowledge graph identification". In: *International Semantic Web Conference*. Springer, pp. 542–557.

Rajpurkar, P., Jia, R., and Liang, P. (July 2018). "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 784–789.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (Nov. 2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392.

Reser, J. P. (2007). "The experience of natural disasters: Psychological perspectives and understandings". In: *International perspectives on natural disasters: Occurrence, mitigation, and consequences*. Springer, pp. 369–384.

Shi, P., Ye, T., Wang, Y., Zhou, T., Xu, W., Du, J., Wang, J., Li, N., Huang, C., Liu, L., et al. (Aug. 2020). "Disaster Risk Science: A Geographical Perspective and a Research Framework". In: *International Journal of Disaster Risk Science* 11.4, pp. 426–440.

Stupina, A. A., Zhukov, E. A., Ezhemanskaya, S. N., Karaseva, M. V., and Korpacheva, L. N. (2016). "Question-answering system". In: *Iop Conference* 155, p. 012024.

Tombu, M. and Jolicoeur, P. (FEB 2003). "A central capacity sharing model of dual-task performance". In: *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-HUMAN PERCEPTION AND PERFORMANCE* 29.1, 3–18.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention Is All You Need". In: *CoRR* abs/1706.03762. arXiv: 1706.03762.

Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014). "Knowledge graph embedding by translating on hyperplanes". In: *Twenty-Eighth AAAI conference on artificial intelligence*.

Weissenborn, D., Wiese, G., and Seiffe, L. (2017). "FastQA: A Simple and Efficient Neural Architecture for Question Answering". In: *CoRR* abs/1703.04816. arXiv: 1703.04816.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *CoRR* abs/1609.08144. arXiv: 1609.08144.

Xiong, C., Zhong, V., and Socher, R. (2016). "Dynamic Coattention Networks For Question Answering". In: *CoRR* abs/1611.01604. arXiv: 1611.01604.

Xiong, C., Zhong, V., and Socher, R. (2017). "DCN+: Mixed Objective and Deep Residual Coattention for Question Answering". In: *CoRR* abs/1711.00106. arXiv: 1711.00106.

Ye, T., Shi, P., Wang, J., Liu, L., Fan, Y., and Hu, J. (June 2012). "China's drought disaster risk management: Perspective of severe droughts in 2009–2010". In: *International Journal of Disaster Risk Science* 3.2, pp. 84–97.

Youn, S. J. (APR 2015). "Validity argument for assessing L2 pragmatics in interaction using mixed methods". In: *LANGUAGE TESTING* 32.2, 199–225.

Zhang, J., Zhu, X.-D., Chen, Q., Dai, L.-R., Wei, S., and Jiang, H. (2017). "Exploring Question Understanding and Adaptation in Neural-Network-Based Question Answering". In: *CoRR* abs/1703.04617. arXiv: 1703.04617.

Zou, M. and Yuan, Y. (Mar. 2010). "China's comprehensive disaster reduction". In: *International Journal of Disaster Risk Science* 1.1, pp. 24–32.