International Journal for Modern Trends in Science and Technology, *8*(09): 290-294, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0809054
Available online at: http://www.ijmtst.com/vol8issue09.html

A Multilayer CNN based approach for Question Answering System

Devina Chaitanya[1], Dr. G.N.V.G. Sirisha[2]

[1]Department of Information Technology, Sagi Rama Krishna Raju Engineering College, Bhimavaram, AP, India.
[2]Associate professor, Department of Information Technology, Sagi Rama Krishna Raju Engineering College, Bhimavaram, AP, India.

To Cite this Article
Devina Chaitanya and Dr. G.N.V.G. Sirisha. A Multilayer CNN based approach for Question Answering System. International Journal for Modern Trends in Science and Technology 2022, 8(09), pp. 290-294. https://doi.org/10.46501/IJMTST0809054

Article Info
Received: 02 September 2022; Accepted: 20 September 2022; Published: 25 September 2022.

ABSTRACT

Question Answering System (QAS) main purpose is to train machines to understand the text like a human and it is one of the challenging tasks in artificial intelligence. QAS is to answer questions based on natural language text. Question Answering System can be viewed as a task of reading a passage of text written in machine understandable language, understanding it, and answering the question based on the corresponding context. The development of deep learning techniques in addition to the availability of large datasets has made the QAS very successful. Due to several advantages in various applications, this QAS system has been used. QAS is applied on several benchmarking datasets such as (MSMACRO, SQuAD, NewsQA, etc.). In this paper, an Improved Question Answering System (IQAS) is introduced by using Stanford Question Answering Dataset (SQuAD). Experiments show the comparison between the proposed approach and existing popular NLP models. This proposed approach is especially focused on the BIDAF and CNN model.

Keywords: Question Answering System (QAS), Stanford Question Answering Dataset (SQuAD), NLP etc.

1. INTRODUCTION

QAS is a significant task that is used to understand the question and gives the relevant answer to the system. These systems are more capable of interacting with the human and system with each other. Figure 1 shows the search engine giving answers to the questions asked by the users based on natural language. The input is in the form of text and this helps the users with high-quality services. This QAS will provide the exact answer of the specified question.

In general, QAS consists of various heuristic rules with advanced analysis like text tagging with parts of speech, unique class tagging, and recognition of entities for various types of "who" questions. Several ML algorithms are combined with the various advanced pre-processing and feature extraction techniques that read the text very efficiently. The performance of ML algorithms may reduce for large datasets and this requires human effort. Some models ignore the dependencies that are failed to extract accurate information. Previously several QA systems are not utilized in real-time applications. To overcome these limitations, the DL algorithms are used to extract accurate information from the real-time datasets. In this paper, Multi-layered CNN is used to

extract the information from the large and real-time datasets. This system solves the issues in DNN architectures and performs better than existing approaches. The proposed approach mainly focused on developing the QAS which is applied to the SQuAD dataset that consists of thousands of context-based questions. The proposed system CNN is integrated with the BiDAF (bidirectional attention flow) model to encode the questions and gives accurate answers to the given question based on the context paragraph. All the results from this system were measured with the parameters such as F1 score, precision, and Exact Match score.

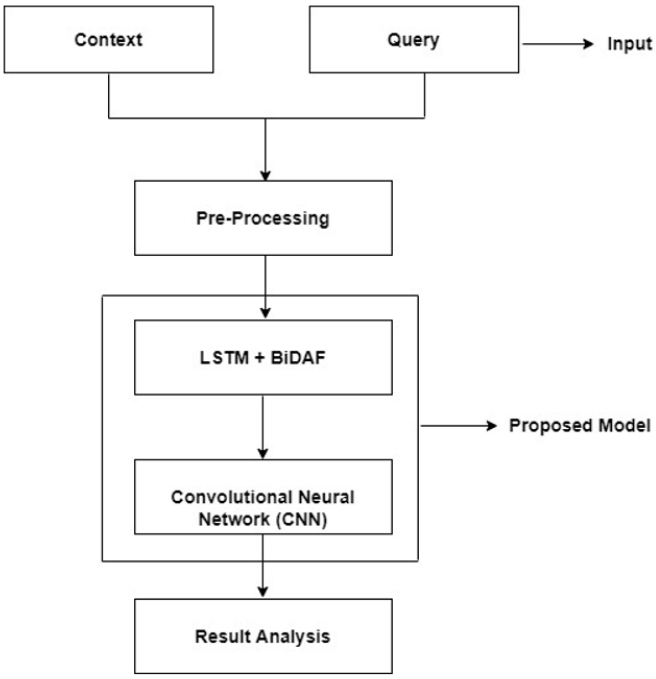

Figure 1: Proposed Architecture

Literature Survey

Wang et al., [1] proposed the advanced neural network system that provides the search for the questions and answers based on user's context. The proposed approach is combined with the matched-LSTM and the textual entailment. This gives the accurate answers for the given text questions. Wang et al., [2] introduced the unique long short-term memory (LSTM) architecture for natural language inference (NLI). This approach mainly focused on significant word-level matching results. The Stanford Natural Language Inference (SNLI) dataset is used for the analysis of accurate results.

Vinyals et al., [3] introduced the neural network architecture that solves the size of variable output dictionaries using a recently proposed mechanism of neural attention. The proposed architecture is called Pointer Net (Ptr-Net). This architecture shows better performance and solves the geometric issues.

Hirschman et al., [4] proposed the automated QAS (A-QAS) which receives the answer very quickly based on the questions given by the user. The automated QAS gives an accurate and valid answer. Presently online search engines give only ranked list documents without giving any answers. This A-QAS addressed and solves these issues and gives better performance.

The Children's Book Test is a piece of bAbI plan of Facebook AI Exploration which targets examining programmed course reading understanding and rationale [5]. Kids books are picked on the grounds that they guarantee an unmistakable story structure which helps this errand. The youngsters stories utilized in CBT come from Project Guntenberg.

Seo et al., [6] Proposed Bi-Directional Attention Flow (BIDAF) which is multi-state hierarchial process that represents the context at different levels of granularity and uses bi-directional attention flow mechanism to obtain a query-aware context representation without early summarization.

Dhingra et al., [7] solve the issues by giving answers to cloze-style questions over documents. The proposed model integrates the multi-hop architecture with the advanced mechanism. This is mainly used to read the documents very effectively.

Kelvin et al., [8] introduced the automated model that extracts the content in the image. This model uses the back-propagation model for training. The performance is analyzed on three datasets such as Flickr8k, Flickr30k, and MS COCO. Weissenborn et al., [9] proposed the simple FastQA which efficiently gives an accurate answer compared with existing approaches. Several issues are identified in the FastQA. To solve this FastQAExt is implemented on several real-time datasets such as SQuAD, NewsQA, and Marco. Compared with FastQA FastQAExt achieved better results.

Kyunghyun et al., [10] proposed the RNN Encoder-Decoder which contains two recurrent models. The first RNN is used to encode the sequence of symbols into fixed-length vector initialization and the second one decodes the initialization to the sequence of symbols. Here, the encoder and decoder model combined gives the maximum training for conditional probability to target the source sequence. Wenhui et al., [11] presented the

gated self-matched networks that show improved question answering that aims to answer the questions from a given input. The input is in the form of a paragraph. The pointer in the proposed model points to the accurate answers from the paragraphs. The experiments are conducted by using the SQuAD dataset. This model achieved an accuracy of 76.1% for the test dataset.

BiDAF model

To find the answer to the question from the given text the following process is used. Firstly, it takes the context and query as an input. Then this model will process the context and query and produces the answer as output from the given context. The following figure shows the methodology of question answering system using BiDAF model.

The entire training of the model will be done in three phases:

1. In the first phase, word embeddings are generated and given as a input to bidirectional RNN.
2. In the second phase, the model trains itself with the context and the corresponding queries using question matching attention flow.
3. In the third phase, the model finds the answer to the question.

Pre-processing

Firstly, it goes through the dataset to separate the context, the questions and answers are placed into separate files. This pre-processing aspect of the dataset was essential to standardize all the data, such as additional symbols or punctuation, and align the indexes of the span with the actual context.

Embedding Layer

Word embedding is used to generate vector representation for the words in the vocabulary. There are two broad categories of word embeddings those are frequency based and prediction based. The GloVe Vectors are used in this project for the word embedding as it combines the best of both worlds. The GloVe vectors are learned by optimizing a loss which uses global count statistics information. This project uses pre-trained GloVe vectors of dimension 100. So, each word in the dataset (context and question) was converted to a vector using these pre-trained GloVe vectors.

Encoding Layer

After the vector representation of words, the next step is to make each word aware of the words before it and after it. This is a very important step as words individually only give partial information. It is the sentence which holds the complete meaning. To make each word aware of its context, it uses a bidirectional Recurrent Neural Network. A recurrent neural network is capable of handling sequential data. The hidden state in the RNN, remembers the previous instances and hence handle the sequential data. They can remember information, which was processed long back, but they face an issue of diminishing gradient which does not allow them to do so. For this reason, this project works with the modification of RNN, Long Short-Term memory (LSTM). LSTM has three gates forget gate, update gate and output gate. Hidden gate is computed using the forget gate and update gate. The output gate determines how much representation the hidden state must have in the output. By the end of this layer, it gives the hidden state vectors for both context and the question.

Attention Layer

The next step would be to understand which part of the context is relevant to the question. This layer is called attention, as it decides where the context to answer the question. This layer will output a vector which assigns weights to each word in the context according to their relevance with the question. Here it is using a Bidirectional attention flow mechanism. It is a high performing attention mechanism. It is based on the idea that attention must flow both ways from question to context as well as context to question. The attention layer takes into question encoding and context encoding and outputs a single vector into which is the attention output.

Output Layer

The main purpose of this layer to predict the span of the answer, from the start index to the end index. With a fully connected layer, the output layer will be able to compute the start index and end index based on the probability distribution vector produced from a SoftMax calculation. With these probability distributions, this system evaluates the basic method, it obtains the maximum index probability from the start and end distribution vectors and returns this as the start index. As a result, this project can maximize the probability through the product of start and end distribution vector to obtain best start and end index. After all these layers have been initialized, the loss layer is used, which simply computes the cross-entropy loss for the start and end index predictions while

also using the Adam optimizer to minimize loss across each batch. The given context and query go through all these layers and process them based on the functionality of the layers. Finally, the result will be the answer to the given question based on the passage will be produced.

Dataset Description

This system uses the SQuAD dataset. This dataset contains 100,000 and more question answer pairs on more than 500 articles. This dataset is split into two parts. Those are training and validation dataset. In terms of dataset splitting, 10% of training dataset is used for validation and the SQuAD 1.1 validation dataset is used for testing. The training dataset contains 23,215 paragraphs with their corresponding five questions and answers pairs and the start index of each answer to the question in the given paragraph.

The CNN model is trained for 2 epochs with a batch size of 8 and each epoch is lasted for 1 hour. The trained model is saved as while training each epoch takes 1 hour of time. So, the trained model is saved in order to save the time. The no of samples that are used to train and validate the model are given below:

Table 1: Total Samples used for Experiments

Total Samples	88539	
Train Samples	80953	91.43%
Validate Samples	7586	8.56%

Performance Metrics

The confusion matrix is used to analyze the performance of proposed and existing approaches by measuring the following metrics.

F1-Score: The precision is the ratio of the number of words in the ground truth to the number of words in the predicted answer. From the overall positives, the percentages of predicted positives are measured. The F1 score is calculated as follows.

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}$$

The precision is calculated as follows. Here tp stands for true positive, fp stands for false positive and fn stands for false negative.

$$\text{Precision} = \frac{\text{tp}}{(\text{tp} + \text{fp})}$$

- tp means the number of words that match with the ground truth and predicted answer.
- fp means the number of words that are in the predicted answer but not in the ground truth.
- fn means the number of words that are in the ground truth but not in the predicted answer.

Accurate Match

This metric is calculated as if the characters of the predicted answer matches with the characters of the ground truth, then the exact match will be 1 otherwise the exact match will be 0. If any single character doesn't match with the ground truth, then the exact match will be 0.

Table 1 shows the performance of Bert and QAS

	Precision	Recall	F1-Score
Bert	68.98	69.78	72.23
CNN	93.56	97.67	93.12

Result Analysis:

Results Using BERT:

The following figures shows the output for Question and Answering System using BERT.

Figure 2: SQuAD Dataset after Training

Figure 3: Output for BERT

Results Using Multilayer CNN:

The following figures shows the output for Question and Answering System using Multilayer CNN.

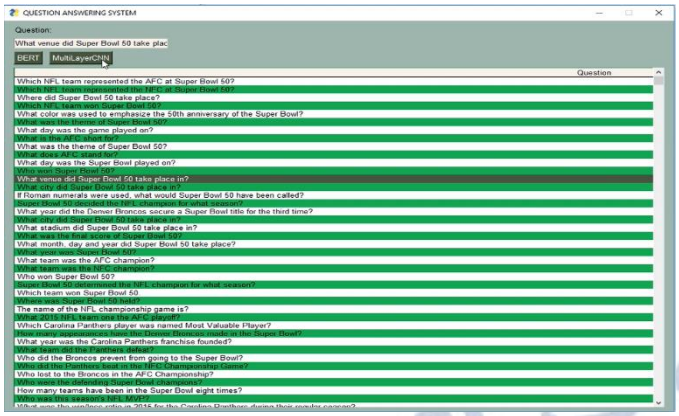

Figure 4: SQuAD Dataset after Training

Figure 5: Output for Multilayer CNN

Performance of BERT And Multilayer CNN models.

MODEL	ACCURACY	F1 SCORE
BERT	69%	51%
Multilayer CNN	85%	67%

CONCLUSION

By using this, the user will get the exact answer to the given question. The answer will be short and precise. By using this system, the human effort is minimized. For loading the data and giving the input, this project will use the squad dataset. The implementation is done by using the CNN and Bert models and comparing both the models to find the best accuracy among them. The performance of the CNN model is improved by training the data on longer sequences and larger batch sizes. By adding linear layer to the model, it will be accurate when compared to previous models.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Wang, Shuohang, and Jing Jiang. "Machine comprehension using match-lstm and answer pointer." arXiv preprint arXiv:1608.07905 (2016).

[2] ang, Shuohang, and Jing Jiang. "Learning natural language inference with LSTM." arXiv preprint arXiv:1512.08849 (2015).

[3] inyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks." Advances in neural information processing systems 28 (2015).

[4] Hirschman, Lynette, and Robert Gaizauskas. "Natural language question answering: the view from here." natural language engineering 7.4 (2001): 275-300.

[5] Hill, Felix, et al. "The goldilocks principle: Reading children's books with explicit memory representations." arXiv preprint arXiv:1511.02301 (2015).

[6] Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." arXiv preprint arXiv:1611.01603 (2016).

[7] Dhingra, Bhuwan, et al. "Gated-attention readers for text comprehension." arXiv preprint arXiv:1606.01549 (2016).

[8] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. PMLR, 2015.

[9] Weissenborn, D., G. Wiese, and L. Seiffe. "Fastqa: A simple and efficient neural architecture for question answering. CoRR, abs/1703.04816." arXiv preprint arXiv:1703.04816 (2017).

[10] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[11] Wang, Wenhui, et al. "Gated self-matching networks for reading comprehension and question answering." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.