The Implementation of Question Answer System Using Deep Learning

Vaishali Fulmal[1], K. P. Moholkar[2], S. H. Patil[3]

[1,2]JSPM's Rajarshi Shahu College of Engineering, Pune, India
[3]Bharti Vidyapeeth College of Engineering, Pune, India
[1]vaishali.fulmali@gmail.com, [2]kpmohalkar_comp@jspmrscoe.edu.in, [3]suhas_patil@yahoo.com

Abstract: Question-answer systems are referred to as advanced systems that can be used to provide answers to the questions which are asked by the user. The typical problem in natural language processing is automatic question-answering. The question-answering is aiming at designing systems that can automatically answer a question, in the same way as a human can find answers to questions. Community question answering (CQA) services are becoming popular over the past few years. It allows the members of the community to post as well as answer the questions. It helps users to get information from a comprehensive set of questions that are well answered. In the proposed system, a deep learning-based model is used for the automatic answering of the user's questions. First, the questions from the dataset are embedded. The deep neural network is trained to find the similarity between questions. The best answer for each question is found as the one with the highest similarity score. The purpose of the proposed system is to design a model that helps to get the answer of a question automatically. The proposed system uses a hierarchical clustering algorithm for clustering the questions.

Keywords: deep learning, LSTM, recurrent neural net- work, automatic question answering, social networking.

1. Introduction

In the area of natural language processing, one of the important challenge is to find if the two sentences convey the same meaning. The question similarity technique can be used for question answering (QA) system [1]. When a user asks a question, the QA system looks for possible similar questions in the available questions. After that, the system identifies the most similar questions using a deep learning algorithm. Answers of the questions which are identified in the previous steps are the correct answer of the asked question. In the question-answer system the important task is to determine the similarity between the questions pair. Answer selection is the task of giving an answer to the existing question which is most similar to the user's question. Recently, the use of machine learning algorithms has increased because these algorithms are capable to solve many difficult tasks in different areas like science and engineering. Deep learning is part of machine learning. Deep learning algorithms can be very efficient in automation of complex tasks. Deep learning methods produce a good performance which is not relying on any feature engineering or expensive external resources. The proposed system provides a new approach in finding similar questions that are available in the dataset relevant to input question. The system is designed with Bi-LSTM and LSTM neural networks and the performance of both the algorithms are compared.

2. Literature Survey

Ziye Zhu (2018) proposes a QA system to select the best answer among the multiple candidates of a question. The authors calculated the similarity using UB-CQA and text classification techniques. The authors uses the information of the user attribute from the response provider. By using this information the system extracts the best answer for a question.

In community-based question-answer system, the user who asks the question has to wait to get the answer and the users who is capable of giving answer to a question has to look for similar question if anyone has asked. Xiang Cheng [5] designs a system that route the questions to the acceptable answers.

M. Breja(2017) designs a QA system that can work in a similar way as the social networking sites. This system provides a platform where the users of the same interest can interact with each other.

G. Zhou (2016) proposes a system which uses a statistical technique to enhance question retrieval as well as

question representation with the translated words from other languages. It uses matrix factorization for this process.

In community question answering systems the answers posted by a user is influenced with the answers which has already been posted by other users for the same question. With time the quality of the answer improves and this process is known as temporal interaction. Causal influence is if the question and answer are appropriate to each other. Fei Wu [2] use both these techniques and LSTM algorithm to find the best answer to the question.

Viriyadamrongkij (2017) proposes a system that finds the difficult questions and distinguishes it from easy ones. Hierarchical technique is used to measure the difficulty level of a question.

J. Liu (2017) proposes a model to distinguish the quality of a question and provide answer using mutual reinforcement. Certain factors which affect the quality of a question are taken into consideration like category, asker, and answer related features to determine the quality.

J. Wang (2016) proposes an online system which can assist the normal medical system to give a relevant answer to the user. The dataset is customized using information obtained from online medical QA sites. The results show improved performance in answer recommendation.

B. Ojokoh (2016) proposes a system where the extraction of quality questions and their classification is done on various academic blogs and websites. The Naive Bayes classification is used in the system.

K.P.Moholkar (2019) proposes a system to identify the relevant answers for a question. Convolutional Neural Network (CNN) is used to extract the features. LSTM algorithm is used to identify the long term dependencies and the context of questions.

3. Proposed System

Relevant dataset of questions from quora is used and the answers are extracted from the internet. In this way, the customized dataset is prepared. The questions are clustered into multiple groups using hierarchical clustering. In general, the merges and splits are determined in a greedy manner in this type of clustering. Bidirectional LSTM is used to retain the context of the questions. The output of BiLSTM is given to a fully connected layer to classify similar types of questions. When a user asks a question, the system identifies the domains, and the system matches the input question with the available questions in the identified domains and calculates the similarity. The system identifies the relevant questions based on the highest matching percentage and provides answers to relevant questions as an output.

The dataset is clustered using hierarchical clustering. The training set is embedded in the system. The bidirectional LSTM is used to retain the context of the question. The output of LSTM is given to the dense layer for the classification of similar types of questions. Figure 1 shows the training phase of the system. The testing phase of the system is shown in figure 2. The trained model is loaded in the system. The user question is the input to the cluster identification block. After identification of the cluster, the model predicts the similarity with the existing questions in the identified domains. The answer is fetched and provided as an output to the user.
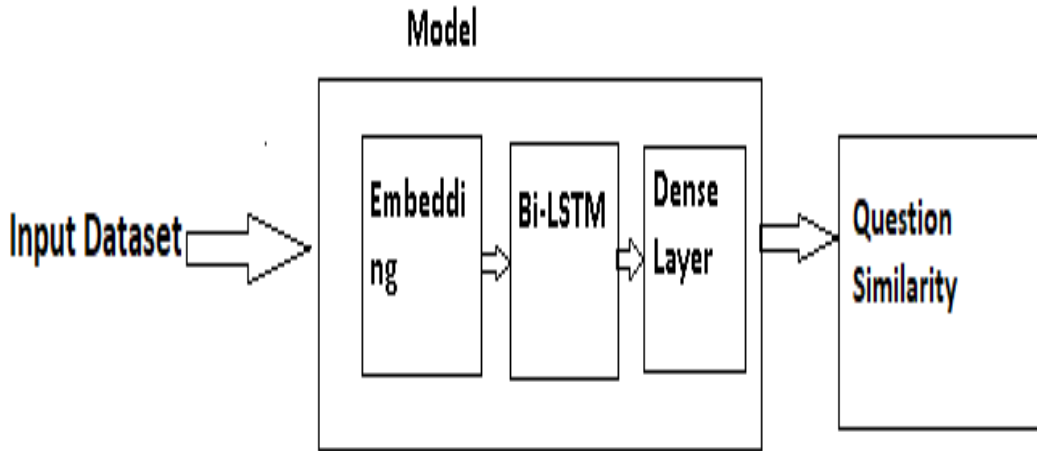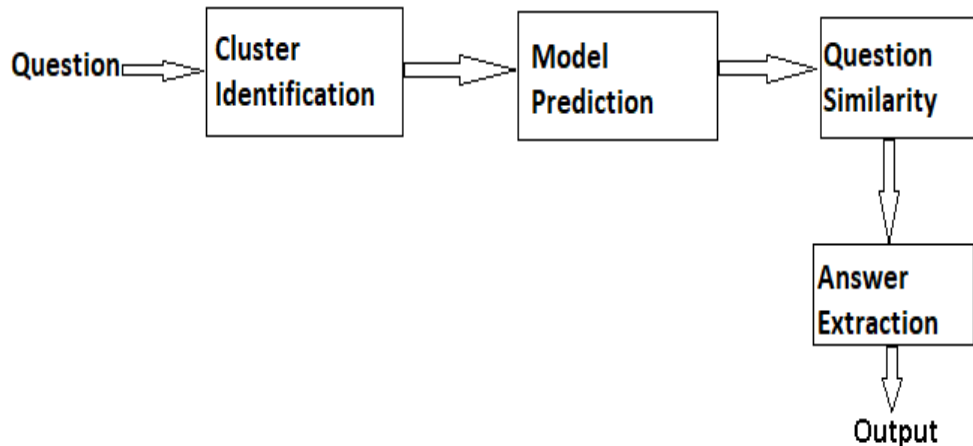
Figure 1. Training the system

Figure 2. Testing the system

4. Algorithm

RNN (recurrent neural network) start reading any document from left and move towards right and after processing each word it updates the state. Problem with traditional RNN is that it losses information about initial words when it reaches to the end of the document. So to retain the state of each words LSTM algorithm can be useful. To have a better result stop words (a, the etc.) should not be considered for training the model, so that the model should not keep any information related to stop words. Selectively read the information added by previous sentiments bearing words (awesome, amazing etc.) and store new information from the current word in the state. This can be achieved using LSTM (Long short term memory) neural networks or Bi-LSTM (for better accuracy). LSTM are type of RNN but it has a capability to store long term information as well. The Bi-LSTM are similar to LSTM but the output in case of Bi-LSTM is connected to the previous state as well as next state. So references of past and future is used in Bi-LSTM neural network and it is more accurate compared to LSTM. The fig 3 shows the basic structure of BiLSTM network.
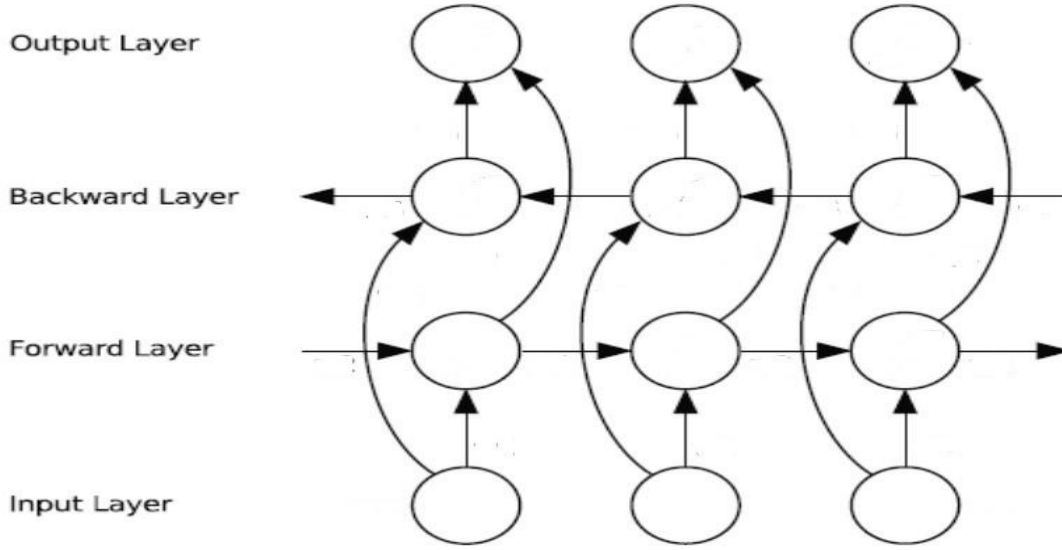
Figure 3. BiLSTM

The LSTM has three gates:

A. Forget gate layer:

This applied to the input at the current time step t and the hidden state at the previous step, i.e. z(t) = [xt, ht−1]. Since the output is a number between 0 and 1 for each element, it controls the amount of information to be retained from the previous time t-1

$$z(t) = (h_{t-1}, x_t) \qquad (1)$$
$$f(t) = \sigma(W_f z_t) \qquad (2)$$

B. Input gate layer:

It is similar to the forget gate, controls which elements of the state vector C have to be updated

$$I(t) = \sigma(W_f z_t) \qquad (3)$$

With these functions, the state C is updated according to the following formula:

$$C_t = f_t * C_{t-1} + I_t * \tilde{C}_t \qquad (5)$$

In other words, the state at the time t depends on the state at the previous time t-1, and by the "important" information that is presented at the time t.

C. Output gate layer:

Finally, the hidden state at the time t is computed, and output is provided if t is also the final time step (i.e. the last element of the input vector)

$$O_t = \sigma(W_O z_t) \qquad (6)$$
$$h_t = O_t * tanh(C_t) \qquad (7)$$

3. Result and Discussion

In the proposed system the quora question pair dataset is used for training. Relevant questions are selected based on the matching percentage from the selected domain's questions. In the pre processing phase the questions are lowercased as well as tokenized to reduce the size. The maximum length of the question is taken as the maximum input size. The word embedding matrix have glove size of 200. Batch size of 100 is used. The learning rate of 1.25

is used. Hidden layers used in this model is 50. Comparison between LSTM and BiLSTM results are shown in below table. BiLSTM gives more accurate results as compared to the LSTM.

Table 1. Results

Modeling Algorithm	LSTM	BILSTM
Accuracy(approx.)	0.80	0.82

4. Conclusion

The system provides satisfactory results in finding the similarity of questions. In this paper, relevant questions of a user's question are obtained using deep LSTM and BiLSTM neural networks, and relevant answers can be fetched and provided as output. Experimental results show that Bi LSTM gives more accuracy as compared to LSTM for finding questions.
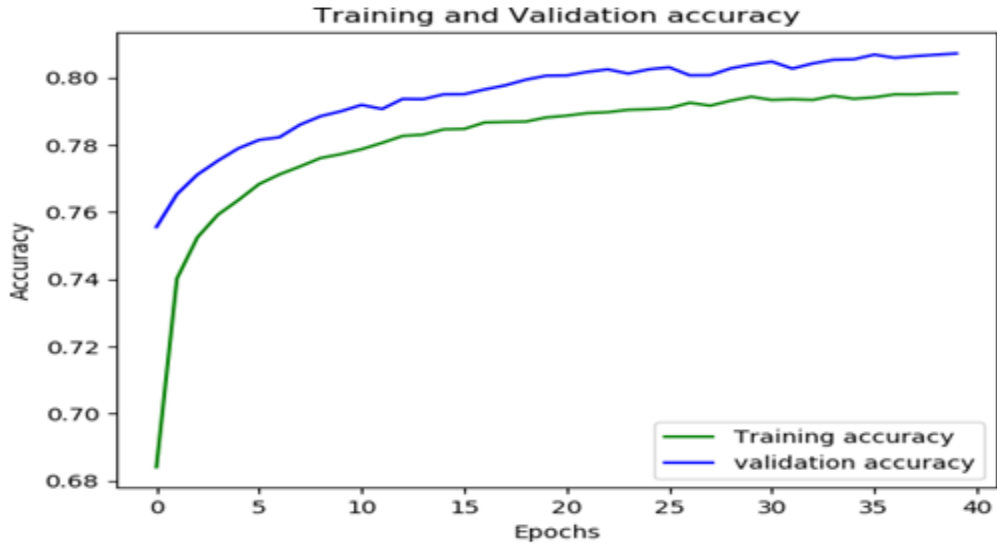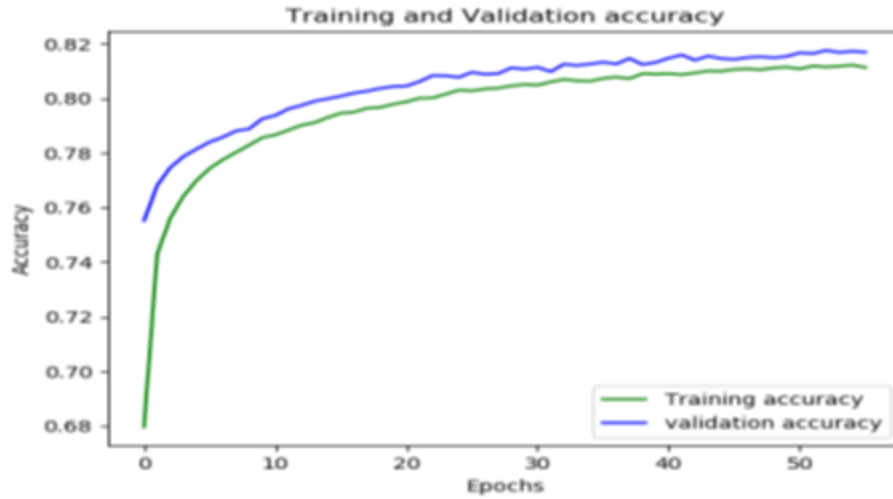

Figure 4. Accuracy for LSTM
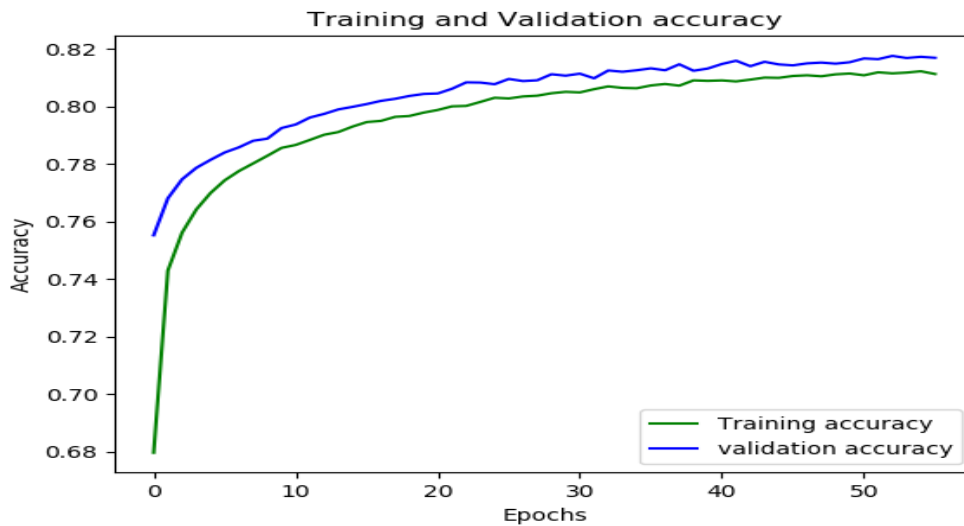
Figure 5. Accuracy for BiLSTM

Figure 6. Loss for Bi LSTM

References

1. Karimi, B. Majidi and M. T. Manzuri,(2019) "Relevant Question Answering in Community Based Networks Using Deep LSTM Neural Networks," 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), Bojnord, Iran, 2019, pp. 1-5.
2. F.Wu et al.(2017), "Temporal Interaction and Causal Influence in Community- Based Question Answering," in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 10, pp. 2304-2317, 1 Oct. 2017.
3. N. Viriyadamrongkij and T. Senivongse (2017), "Measuring difficulty levels of JavaScript questions in Question-Answer Community based on concept hierarchy," 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE), Nakhon Si Thammarat, 2017, pp. 1-6.
4. J. Liu, H. Shen and L. Yu (2017), "Question Quality Analysis and Prediction in Community Question

Answering Services with Coupled Mutual Reinforcement," in IEEE Transactions on Services Computing, vol. 10, no. 2, pp. 286-301, 1 March-April 2017.

5. X. Cheng, S. Zhu, S. Su and G. Chen, (2018), "A Multi-Objective Optimization Approach for Question Routing in Community Question Answering Services (Extended Abstract)," 2018 IEEE 34th International Conference on Data Engineering (ICDE), Paris, 2018, pp. 1765-1766.

6. M. Breja,(2017) "Social network analysis in question answering community," 2017 International Conference on Energy, Communication, Data Ana- lytics and Soft Computing (ICECDS), Chennai, 2017, pp. 314-318.

7. G. Zhou, Z. Xie, T. He, J. Zhao and X. T. Hu(2016), "Learning the Multilingual Translation Representations for Question Retrieval in Community Ques- tion Answering via Non-Negative Matrix Factorization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1305-1314, July 2016.

8. J. Wang, C. Man, Y. Zhao and F. Wang. (2016), "An answer recommendation algorithm for medical community question answering systems," 2016 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), Beijing, 2016, pp. 139-144.

9. B. Ojokoh, T. Igbe, A. Araoye and F. Ameh. (2016), "Question identification and classification on an academic question answering site," 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), Newark, NJ, 2016, pp. 223-224.

10. Z. Zhu, X. Liu, H. Li and T. Li. (2017), "UB-CQA: A user attribute based community question answering system," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017.

11. Moholkar, Kavita P. and Suhas Haribhau Patil. (2019) "Hybrid CNN-LSTM Model for Answer Identification." 2019.

12. K. P. Moholkar, S.H. Patil(2019), "A Question Answer System: A survey", International Journal of Computer Sciences and Engineering, Vol.7, Issue.3, pp.441-447, 2019

13. Multiple Choice Question Answer System using Ensemble Deep Neural Network – Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020), 5-7 March 2020, Scopus Indexed ISBN : 978-1-7281-4167-1 https://ieeexplore.ieee.org/document/9074855