International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

The Question Answer System Using Deep Learning

Vaishali Fulmali Prof K.P. Moholkar

Department of Computer Engineering JSPM's Rajarshi Shahu College of Engineering Savitribai Phule Pune University Pune,India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Question-answer systems are referred to as intelligent systems that can be used to provide answers for the questions which are asked by the user based on certain facts or rules stored in the knowledge base. 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 were becoming popular over the past few years. It allows the members of the community to post as well as answer the questions. It enables general users to seek information from a comprehensive set of questions that are well-answered. CQA texts are relatively noisy and therefore the implementation of the classic text mining methods such as a bag of a word doesn't lead to good results. In the proposed system, a deep learningbased model is used for automatic question-answering. First, the questions and answers 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 designing an automated question answering system. The proposed system uses a hierarchically clustered dataset. This dataset is the basis for the QA system. Finally, this system can provide a satisfactory answer for the asker.

Keywords: *Social network, Community based question answering, recurrent neural networks, LSTM, deep learning*

Introduction

In the area of natural language processing, one of the important problems is to determine whether two sentences have approximately the same meaning. The important application of question similarity is the Question Answering (QA) system. Upon receiving a question from the user, a QA system first retrieves a list of possible questions from a large dataset, resulting in only a few hundred ones. Then, it selects the most similar question set from the list with the help of a deep learning algorithm. The answers to these existing questions are most likely the correct responses to the original question. In such a system, the key issue is to determine the questionquestion similarity, i.e., find a question that is most similar to the user's question. Answer selection is the task of giving an answer of the existing question which is most similar to the user's question. In recent times, machine learning algorithms can solve many complex tasks in various areas of science and engineering. Deep learning which is part of machine learning provides algorithms giving high accuracy solutions for automating various tasks. Deep learning methods produce a good performance which is not relying on any feature engineering or expensive external resources. In the proposed system, a novel method for finding relevant questions to the question of a user, using LSTM based neural networks is used.

Literature Review

Ziye Zhu [21] proposes a user attribute based community question answering system in which the problems like how to make a choice of the best answer among multiple candidates for a single question are taken into consideration and the user attribute based community question answering system is designed. First, in database refinement part , authors construct an enhanced and organised database by the best answer choosing method leveraging user attribute provided by the response provider. Second, in human-computer part, the UB-CQA is able to look and provide a more fulfilling answer to users by similarity calculation and re-ranking method leveraging text categorization information. Empirical evaluations show that the answer choosing and the candidate question re-ranking methods bring great improvements in accuracy and reliability. Data analysis shows that authors took full advantage of the user attribute information provided by the response providers to extract the best answer for each question in the original Q-R database and construct an optimized and structured database, which is the basis of the CQA. Fei Wu [2] uses the temporal interactions between answers (how previously posted answers impact the lately posted answers). For example, a rational user usually adapts others opinions and then revises his inclinations, and posts a more appropriate answer after understanding the given question and previously posted answers to the question. The system proposes an architecture which is named Temporal Interaction and Causal Influence LSTM also known as TC-LSTM to effectively take advantage of not only the causal influence between question-answer that is how appropriate an answer is for a given question but also the temporal interconnections between answers-answer that is how a high-quality answer gradually forms. In particular, long short-term memory is used to find the explicit question-answer influence and the implicit answers-answer interactions. Experiments are conducted on SemEval 2015 CQA dataset for answer classification task and baidu zhidao dataset for answer ranking task. The experimental results show the advantage of the model comparing with other state-of-the-art methods.

N. Viriyadamrongkij [3] proposes measuring difficulty levels of javaScript questions in question-answer community which is based on concept hierarchy. The method is used to measure question difficulty levels which are directly based on the question contents. In particular, authors analyzes the difficulty of terms that come into the sight in a JavaScriptrelated question, based on the given JavaScript concept hierarchy. In the judgement of the performance of the extent of question difficulty, concept based measure gives alike performance to that of the real estimations based on the features of the QAC. When they both are used together, the performance can be improved.

J. Liu [4] proposes question quality analysis and prediction in community question answering services with coupled mutual reinforcement. The existing works analysis shows how asker and question-related features affect the Question Quality (QQ) related to the amount of attention from users and the number of answers and the question solving latency, but neglect the relation between QQ and AQ which is measured by the rating of the best answer. The task of finding this relation is critical to quality of service (QoS). Authors handle this problem from two aspects. Initially additional use of QQ in measuring AQ, and analysis of correlation between a comprehensive list of features (including answer-related features) and QQ is done. Then the authors propose the method that estimates the probability for a given question to obtain high AQ. Their analysis on the Yahoo! Answers trace confirmed that the list of identified features exert influence on AQ, which determines QQ. For the analysis of correlation, the classification algorithms cannot consider the mutual interactions between multiple that is greater than two classes of features. Authors then propose a novel coupled semi-supervised mutual reinforcement-based label propagation (CSMRLP) algorithm. The CSMRLP(Coupled Semi-Supervised Mutual Reinforcement-based Label Propagation) outperforms the Mutual Reinforcement-based Label Propagation (MRLP) and five other traditional classification algorithms in the accuracy of AQ classification, and the effectiveness of their proposed method in AQ prediction. Finally, they provide suggestions which can be used to derive full benefit for improvement of the QoS of CQAS.

J. Wang [8] proposes an answer recommendation algorithm for medical community question answering system. The system introduces the experiments to evaluate the algorithm on the datasets which are collected from the online medical query answering system named ask.39.net. Next, the results demonstrate the effectiveness of the proposed methods. The experiments contain three parts, the similar cases retrieval, answer quality estimation and answer recommendation. The algorithm used in this system is tested on a dataset collected from online medical community question answering system. The results shows improved performance in answer recommendation.

B. Ojokoh [10] proposes question identification and classification on an academic question answering site. The system consist of systematic approach to identification and classification of questions. Semantic appearance of part of speech tag in english language is used to identify the questions. The Classification is done based on maximum probability value of Na¨ıve Bayes classification. The model is validated and assessed with experiments on some crawled web pages from ResearchGate.

Proposed System

The questions are extracted from Quora dataset. Corresponding answers are extracted and customized dataset is prepared using best answers to each questions. The questions are clustered using hierarchical clustering. Hierarchical cluster analysis or hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. In general, the merges and splits are determined in a greedy manner in this type of clustering. The outcomes of hierarchical clustering are generally represented in a dendrogram. Bidirectional LSTM is used to retain the context of the quetion and answer. The output of BiLSTM is given to fully connected layer to classify similar type of questions . When user ask a question, the system identifies a domain and then identifies question similarity between existing questions. The system extracts answers which are relevant to the users question. The input dataset is clustered using hierarchical clustering. The training set is embedded in the system.The bidirectional LSTM is used to retain the context of question and answer. The output of lstm is given to the dense layer for classification of similar type of questions. The figure 1 shows the training phase of the system. The testing phase of system is shown in figure 2. The question is input to the cluster identification. After identification of cluster the model is predicted. Identification of the question similarity between the existing questions is done and the answer is extracted.

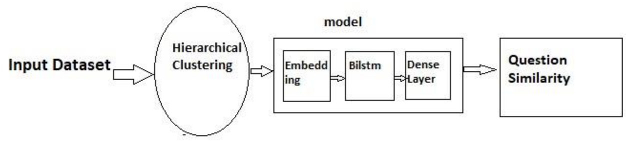

Fig. 1. Training the system

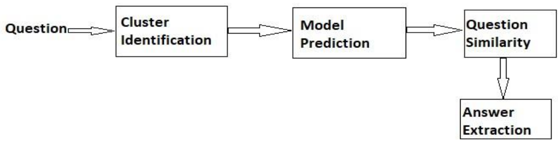

Fig. 2. Testing the system

Algorithm

The LSTM deep neural networks, are a particular type of recurrent neural networks which have the ability to learn long term relationships. Recalling information for long periods of time is the default behavior of ordinary LSTM networks. For Bi-directional-LSTM as shown in fig 3 the output will be connected both to the previous state and to the next state. Therefore, there is the possibility of prediction based on the past and the future information. The neurons are normal LSTM layer divided into two neurons, one for the forward-mode and the other for the backward-mode.

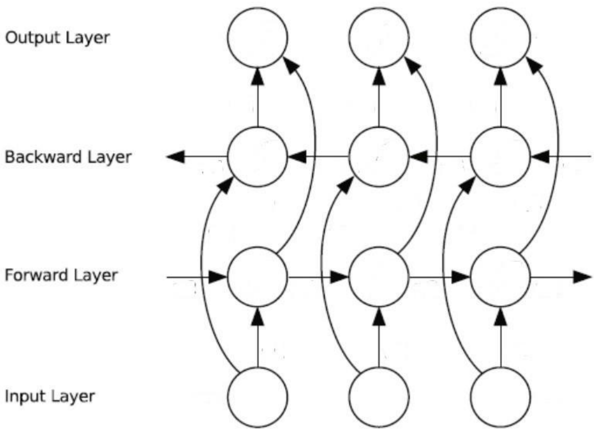

Fig. 3. BiLSTM

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) = [x(t), h(t-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 step t-1

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

B.

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

$$I(t) = \sigma\left(W_f z_t\right)$$
$$\tilde{C}_t = \tanh\left(W_C z_t\right)$$

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

$$C_t = f_t \cdot C_{t-1} + I_t \cdot \tilde{C}_t$$

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

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

$$O_t = \sigma\left(W_O z_t\right)$$
$$h_t = O_t \cdot \tanh(C_t)$$

Result and Discussion

In this work, we have finished with training the dataset.The table shows the training dataset results.
We have used Quora dataset which is applied to both LSTM and BiLSTM and the graph in figure 5 shows the number of similar questions recognised.

LSTM

Phase	No. of questions	Domain Recognized	Similar Question Found
1	100	100	29
2	100	100	32
3	100	100	35

BiLSTM

Phase	No. of questions	Domain Recognized	Similar Question Found
1	100	100	35
2	100	100	41
3	100	100	45

Fig. 4. Training Dataset Results

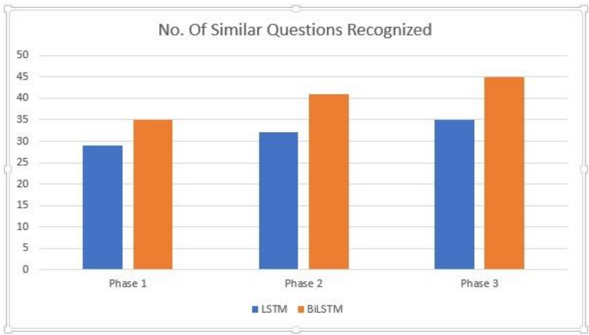

Fig. 5. Number of similar questions recognised

Conclusion

The system provides satisfactory results in finding similarity of question. It extract the best answer for each question in the dataset, which is the basis of the CQA. Best answer choosing method significantly improves the quality of the answers for users.

References

[1]. E. Karimi, B. Majidi and M. T. Manzuri, "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., "Temporal Interaction and Causal Influence in CommunityBased 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, "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, "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, "A Multi-Objective OptimizationApproach 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, "Social network analysis in question answering community," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 314-318.2016 International Conference on Wireless Communications, SignalProcessing and Networking (WiSPNET), Chennai, 2016, pp. 2473-2477.

[7]. G. Zhou, Z. Xie, T. He, J. Zhao and X. T. Hu, "Learning the Multilingual Translation Representations for Question Retrieval in Community Question 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]. [8] J. Wang, C. Man, Y. Zhao and F. Wang, "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]. [9] M. R. Sumalatha and N. A. Priyanka, "Analyzing and predicting knowledge of contributors in community question answering services,"

[10]. B. Ojokoh, T. Igbe, A. Araoye and F. Ameh, "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.

[11]. J. Cao, F. Jay, and J. Nunamaker,"Question answering on lecture videos: A multifaceted approach",in Proc. Int. Joint Conf. Digital Libraries, 2004.

[12]. L. A. Adamic, J. Zhang, E. Bakshy, andM. S. Ackerman, "Knowledge sharing and Yahoo answers: Everyone knows something"in Proc. Int. WorldWide Web Conf., 2008.

[13]. T. Yeh, J. J. Lee, and T. Darrell, "Photo-based question answering", in Proc. ACM Int. Conf. Multimedia, 2008.

[14]. S. A. Quarteroni and S. Manandhar, "Designing an interactive open domain question answering system", J. Natural Lang. Eng., vol. 15, no. 1, pp. 73-95, 2008.

[15]. R. C. Wang, N. Schlaefer, W. W. Cohen, and E. Nyberg, "Automatic set expansion for list question answering", in Proc. Int. Conf. Empirical Methods in Natural Language Processing, 2008.

[16]. Y.-C. Wu and J.-C. Yang, "A robust passage retrieval algorithm for video question answering", IEEE Trans. Circuits Syst. Video Technol., vol. 18,no. 10, pp. 1411-1421, 2008.

[17]. G. Li, H. Li, Z. Ming, R. Hong, S. Tang, and T.-S. Chua, "Question answering over community contributed web video", IEEE Multimedia, vol. 17,no. 4, pp. 46-57, 2010.

[18]. M. Wang and X. S. Hua, "Active learning in multimedia annotation and retrieval: A survey", ACM Trans. Intell. Syst. Technol., vol. 2, no. 2, pp. 10-31, 2011.

[19]. Liqiang Nie, Meng Wang, Zheng-Jun Zha, Guangda Li and Tat-Seng Chua"Multimedia Answering: Enriching Text QA with Media Information"July 24–28, 2011.

[20]. Liqiang Nie, MengWang,Yue Gao, Zheng-Jun Zha and Tat-Seng Chua,"Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information", VOL:15 NO:2 in 2014.

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

[22]. Moholkar, Kavita P. and Suhas Haribhau Patil. "Hybrid CNN-LSTM Model for Answer Identification." 2019.