Turkish Journal of Computer and Mathematics Education Vol.12 No.10 (2021), 3168-3172

Research Article

Student Suite+, A Closed Domain Question Answering System for Educational Domain

V.Swathilakshmi[a], Rama Satyanarayanamma M[b], Vismaya Udayakumar[c], and Satyavarapu Sai Sindhura[d]

[a]Assistant Professor, Department of Computer Science & Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India
[b,c,d]Department of Computer Science & Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI). It is a study of how machines understand the language of humans. It aims to build systems that can understand text and perform tasks like language translation and classifying a topic. NLP has manyapplications. NLP tools help to process unstructured data. Question

Answering (QA) system or information retrieval system falls under the category of NLP. The QA system mainly involves two methods, question analysis andanswer extraction. Intelligent QA systems have always outperformed the non-intelligent QA systems. It is a neural network that is incorporated with the QA system to make it intelligent. Neural networks(NN) are a series of algorithms that are used to mimic the operations of a human brain to understand the relationships between data. NN can learn by themselves and produce the output that is not limited to the input provided to them.

Keywords: NLP, Natural Language Processing, Question Answering (QA) system, Neural Networks, Closed Domain QA system.

1. INTRODUCTION

NLP is a branch of AI that assists computers to understand, interpret and manipulate the languages spoken by humans. NLP acts as the bridge betweenhumans and computers. NLP helps computers communicate with humans in the selected language. For example, NLP makes it possible for computers to read text, hear speech, analyse, understand the sentiment, and determine which parts of the text or speech are important. Today's machines can process more data than humans, without weariness and steadily, in an unbiased way. Huge amounts of unstructured data is generated every second. We humans have more than 100 languages out of which some are very complex and not so easy to understand. Humans express our emotions and feelings in many ways, both verbally and in writing. There are hundreds of languagesand scripts, but within each language is a unique setof grammar and syntax rules, terms and slang. When we write, we often tend to make mistakes. When we speak, we have regional accents, also some bilinguals tend to borrow terms from other languages. NLP is a crucial element because it helps resolve issues in language and adds useful numeric structure to the data for many applications. The major steps involved in NLP is accepting inputfrom the user, processing the input and extracting the entities. These major modules are further divided as provided in the image.

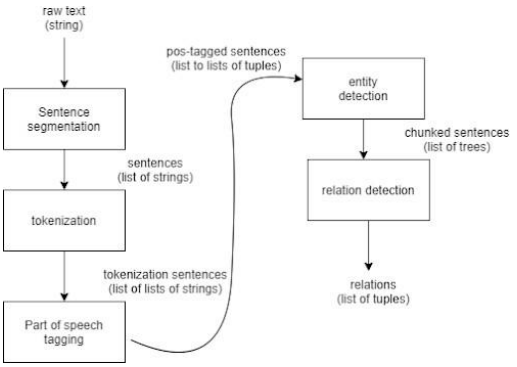

Figure 1.Steps involved in NLP

2. RELATED WORKS

Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automaticallyanswer questions posed by humans in a natural language. A QA System, usually a computer program, may construct its answers by querying a structured database of knowledge or information, usually a knowledge base. More commonly, question

answering systems can pull answers from an unstructured collection of natural language documents. Neural networks, also known as artificial neural networks or simulated neural networks are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Question Answering systems have been implemented even before the concept of neural networks were used. They were either knowledge base related or involved experts answering the questions. A lot of questions are being asked in forums and the web. When the number of questionsasked were low, it was an easy task for the experts to find the questions that they can answer. But as the number of questions asked increased, it is not necessary that the questions can be found by the expert without any sorting or forwarding techniques. In (Haiying Shen,2015), the questions were asked by the users and the questions were forwarded to other users who seemed to have the same common interest as of the category of the question. These questions were forwarded to other users using the concept of user interest analyser, question categorizer and question user mapper. Also, the closeness of the user with the expert and the answering frequency of the expert was also calculated. The user interest analyser makes use of the user profile, analyses it, and creates a vector to mention the interests of the user. The question categoriser is used to find the category under which the question comes. For both user interest analyser and question categoriser, word net is used. (Yuhua Lin, 2015) Question user mapper decides which expert userthe question must be forwarded to. It is decided with other factors like closeness of the expert with the user and answering frequency. The issue with this model is that there is no factor to make sure that the expert is providing the right answer and how effective the answer is. And also there is no mechanism to fight the spammers involved in the system. These issues were solved in. They have added a feature of accepting user feedback which allows the system to know whether the answer provided by the expert was understandable and relevant. It is a reputation-based system which is capable of solving the issues involved. This system forwards the questions to the reputed users who are experts in some subject. The reputation is decided based on major factors like willingness to answer, reputation score, direct trust and feedback from other users. The willingness to answer is calculated based on the ratio of number of questions forwarded to the expert and the number of questions answered by the expert. Reputation score is calculated with the help of the feedback received from the users. The spam detection system allows to identify and remove spammers involved in the Network. QA system is the improvised version of search engines or can be stated as specific search engines because it provides answersin the form of URLs whereas QA systems produce direct answers. That is why more focus has been put on developing better QA systems that could respond to a large variety of user query. Semantic Web Technology has aided the development of such systems from past many years because it could add meaning to the query being asked. (Shally Garg, 2016) makes use of semantic web technology along with natural language processing to answer the queriesof the user which will be regarding the programming language java. The architecture consists of 5 main modules : Query Analyzer, Keyword Extractor, Source Ontology, Query reformulation, Query Processing. and can beused to provide factoid questions. In [9], the first step is finding the type of question, whether it is a yes or no question or a one word answer questionor a question that will require sentences to provide the correct answer. The ambiguity in the question is also removed to avoid any further problems. The top results from google are fetched. (Prakash Ranjan, 2016) All these methods do not involve any sort of memory element or intelligence which would have been an added advantage. As technologies like neural networks were found, neural networks allow the computer to think and make decisions like a human. The involvement of neural networks in the field of question answering has led to a lot of different new methods of question analysis and answer mapping. As the concept of neural networks such as DNN, CNN, RNN and its variants like GRU, LSTM and many more, Question answering systems are being researched heavily with another added advantage of the advancement in the field of NLP. In (Wei-Nan Zhang, 2016) neural networks have been used for query reinforcement using rank-based methods to vary the weights of the words or elements using which they are enhancing the question. It is used for key concept identification which helps in capturing the semantics. This method is used to extract the features from the question raised by the user in order to understand what the user is trying to getthe answer for. In previous papers there had occurred a lot of issues regarding word mismatch and verboseness. These issues are solved in this system using this approach. In (Hasangi Kahaduwa, 2017) the task to be performed is divided into 2 parts, question identification and answer search in the knowledge base. Here, the knowledge base is regarding the travel domain. So, this falls under the category of closed domain question answering system. Rule based approach has been used in whereas rank-based method was used. A machine learning approach has been used to identify the question with the help of syntactic and semantic features. Answer retriever from the knowledge baseis performed using a rule based approach. This approach involves filtering the answers from the knowledge base to provide the most relevant answer. Dynamic memory network (DMN), a neural network architecture which processes input sequences and questions, forms episodic memories,and generates relevant answers is the concept used in (Lei Su 2019)]. The memory element allows us to enhance the quality of the answers provided. The added memory and reasoning ability also plays a major role in answer selection. Entity linking is the task of assigning a unique identity to entities mentioned in text. Using entity linking, the answer selection process can be simplified. Then

the question answer pair is sent to the dynamic memory networkfor a final answer.

As per mentioned in, it has the best matching aggregation framework and is capable of giving answers with higher accuracy while compared to other Knowledge Base Question Answering (KBQA) systems. It provides an optimal matching result between questions and candidate answers based on sequential matching and by choosing the most useful context of candidate. They have incorporated question-specific contextual relations connected to a candidate to enhance its representation, where we use attention mechanismsto weigh the relations based on their relevance to the question . Whereas in the weights of the words or elements are used to enhance the question. Even though the model proposed in performs better than the rest of the KBQA it still cannot work with complex questions. The questions asked by the user will be processed by the question answering system. The input provided will be in the form of text or speech. If it is in the form of speech, it has to be converted to text beforethe analysis begins.

A lot of approaches have been implemented and suggested for the smooth conversion of speech to text. It deals with the conversion of emotional speech to text. Emotional speech is a little difficult to convert to text as the person might be talking in a higher pitch or even elongate the words. Under such conditions, a common system might not be able to understand properly. The system proposed in can work efficiently under such situations. Even though the accuracy is not very remarkable the system was capable of providing an accuracy of 75%. In , the proposed system can extract the speech even though it was delivered in a noisy environment. The RNN used has also played a major role in providing an accuracy of 74 - 75%. As per our understanding, while comparing all the existing approaches, it is obvious that there are some sort of issues related to question analysis or answering. It is capable of providing better results than any of the above discussed papers. It cannot provide answers when complex questions are asked but the approach used in allows it to extract the entities in the question easily and provide the most relevant answer using the transformer based neural network which has Bidirectional Long Short Term Memory algorithm (BiLSTM) as the backbone. Transformer based feature extractor is used for question extraction. This approach is implemented for the dataset WikiQA. WikiQA is a publicly available set of question and sentence pairs, collected and annotated for research on open domain question answering. WikiQA is constructed using a more natural process. Eventhough It is capable of providing answers, there are some situations where it cannot provide the right answers. State-of-the-art machine learning algorithms seem to provide better results. The same approach can be applied on closed domain question answering systems using different data sets and is also capable of providing better answers.

3. PROPOSED SYSTEM

The proposed system is a closed domain question answering system focusing mainly on the educational domain. A question answering system accepts questions from the user, extracts the entities from the input, searches for answers in the databaseand provides the most relevant answer. Thisquestion answering system provides answers for all questions related to the topics provided in the NCERT text books. One of the major drawbacks of the existing question answering systems is that they are not capable of providing answers according to the level of understanding of the student. The proposed QA system will first find the class in which the student is studying in and then answer his or her questions according to the content in their textbook. This will be one of the major advantages of this model over the existing question answering systems available.

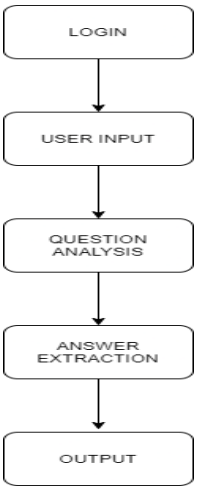

Figure 2.Architecture

3.1 Login

This is one of the most basic modules. Only the class selection happens in this module. As soon as the user has opened the application, they will be asked to enter the class that they are currently enrolled in, this allows it to load the required model from the range of trained models available. When features like profiles are added, separate login with authentication will be included. If the user wishes to view the contents of another class, he or she can go back to the login part and change the class.

3.2 User Input

Any information or data sent to a computer for processing is considered input. Input or user input is sent to a computer using an input device. The proposed model is a question answering system capable of receiving queries from the student and providing answers accordingly. So, it is necessary that it gets an input from the user to provide answers. The question answering system will accept inputs in 2 forms, text and audio. Input given in text format can be processed easily and head to the next step. But it is not the same for audio input. The audio is processed to extract the message out of it. This process is known as speech recognition, the process of extracting words or sentences from audio. Now that the message or query is extracted, it can move to the next step or next module.

3.3 Question Analysis

After extracting the query successfully, the next step is to find what the user wants to know. The text is thoroughly analysed to find what the user is trying to know. The answer to ‘ When is Independence Day’ and ‘ What is Independence Day’ is completely different. The question changes completely when ‘When’ is replaced with ‘What’. Finding the main keyword, which is Independence Day’ in this example, and then retrieving a general definition will not be useful. This is the reason why the message is analysed thoroughly to find what exactly the user is trying to find about. The main entities are extracted and analysed to know what kind of answer the user wants to know about.

3.4 Answer Extraction

The question is known and now is the time to find the right answer. As mentioned above, it is necessary to make sure that the best answer is being provided to the user. This step involves finding the appropriate answer. The knowledge base will have a lot of other information and the task is finding the most relevant and best answer available in the knowledge base. Here, we will be using a transformer based model for the process of extracting the answer. The Retriever is a lightweight filter that can quickly go through the full document store and pass a set of candidate documents to the Reader. It is a tool for sifting out the obvious negative cases, saving the Reader from doing more work than it needs to and speeding up the querying process.

3.5 Output

Now that the answer is obtained, it is time we provide the user the same. The information given back to the user after all the steps of various levels of processing is called output. Just like the input,the output can be given in 2 formats, Audio and Text. Output in textual format would not be much of a task. If the output is to be given in audio format, the text will have to be converted to audio, which should sound like someone is reading out the answer to them, just like the different chatbots and assistants like Alexa and Siri does.

4. CONCLUSION

There are a lot of questions answering systems available, open domain and closed domain. Open domain question answering systems are capable of answering questions related to any domain whereas closed domain question answering systems deal with answering questions related to a particular domain. This paper majorly focuses on building a closed domain question answering system for the school going students who follow the NCERT syllabus. One of the main advantages of this model is that it is capable of answering questions according to the class which the student is studying in and also ensures that the student is given an answer in his or her textbook rather than providing them with answers which they might not be able to understand.

This model answers questions related to the topics given in the NCERT textbook. The same model can be used for different syllabuses but will have to be trained with their respective data which is the data in the textbooks. This paper deals with textbooks where the content is provided in English.The same QA system can be enhanced to deal with question and answering in other languages other than English. Another feature that can be added is providing the answers in the language selected by the user. The same application can be connected with the school software and be used to view the details like attendance, fee due, etc. By adding all these features, the application will be of great useto the students.

REFERENCES

1. Cassia Valentini-Botinhao, Junichi Yamagishi, Senior Member (2018) *“Speech Enhancement of Noisy and Reverberant Speech for Text-to-Speech”, IEEE/ACM,*

2. Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, Bruce Croft and Mark Sanderson (2017) *"Document Summarization for Answering Non-Factoid Queries", IEEE*
3. Ezhilarasi T. P, G.Dilip, T.P.Latchoumi, K.Balamurugan* (2020*), UIP—A Smart Web Application to Manage Network Environments, Advances in Intelligent systems and computing book series, https://doi.org/10.1007/978-981-15-1480-7_8, 97-108.*
4. Haiying Shen, Guoxin Liu, Nikhil Vithlani (2015) *"SocialQ&A: An Online Social Network Based Question and Answer System", IEEE*
5. Hasangi Kahaduwa, Dilshan Pathirana, Pathum Liyana, Arachchi, Vishma Dias, Surangika Ranathunga (2017)
6. *"Question Answering System for the Travel Domain", IEEE*
7. Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang and Chengqing Zong (2018) *"Read, Watch, Listen and Summarize: Multi-modal Summarization for Asynchronous Text, Image, Audio and Video", IEEE*
8. Latchoumi T. P, K. Balamurugan, K. Dinesh and T. P. Ezhilarasi, (2019). *Particle swarm optimization approach for water-jet cavitation preening. Measurement, Elsevier, 141,184-189.*
9. Latchoumi T. P, T. P. Ezhilarasi, K. Balamurugan (2019), *Bio-inspired Weighed Quantum Particle Swarm Optimization and Smooth Support Vector Machine ensembles for identification of abnormalities in medical data. SN Applied Sciences (WoS), 1137, 1-12, DOI: 10.1007/s42452-019-1179-8.*
10. Latchoumi, T. P., Reddy, M. S., & Balamurugan (2020), *K. Applied Machine Learning Predictive Analytics to SQL Injection Attack Detection and Prevention. European Journal of Molecular & Clinical Medicine, 7(02), 3543-3553*
11. Lei Su 1, Ting He1, Zhengyu Fan1, Yin Zhang2 and Mohsen Guizani3 (2019) *"Answer Acquisition for Knowledge Base Question Answering Systems Based on Dynamic Memory Network", IEEE Access*
12. Luca Cagliero, Laura Farinetti, and Elena Baralis (2019) *"Recommending Personalized Summaries of Teaching Materials", IEEE Access*
13. Prakash Ranjan, Rakesh Chandra Balabantaray (2016) *"QUESTION ANSWERING SYSTEM FOR FACTOID BASED QUESTION", IEEE*
14. Pruthviraju G, K.Balamurugan*, T.P.Latchoumi, Ramakrishna M (2021), *A Cluster-Profile Comparative Study on Machining AlSi7/63% of SiC hybrid composite using Agglomerative Hierarchical Clustering and K-Means, Silicon, 13, 961–972, DOI: 10.1007/s12633-020-00447-9, Springer.*
15. Rafael Dantas Lero,Dr Chris Exton Lero, Dr Andrew Le Gear (2019) *"Communications using a speech-to-text-to-speech pipeline", IEEE Xplore*
16. Shally Garg, Suresh Kumar (2016*) "JOSN: JAVA Oriented Question-Answering System Combining Semantic Web and Natural language Processing techniques", IEEE*
17. TAIHUA SHAO, YUPU GUO, HONGHUI CHEN, AND ZEPENG HAO (2019) *"Transformer-Based Neural Network for Answer Selection in Question Answering", IEEEAccess*
18. Tae-Ho Kim, Sungjae Choy, Shinkook Choiy, Sejik Park and Soo-Young Lee (2020) *"Emotional voice conversion using multitask learning with text-to-speech", IEEE Xplore*
19. Venkata Pavan M,Balamurugan Karnan*, Latchoumi T.P (2021), *PLA-Cu reinforced composite filament: Preparation and flexural property printed at different machining conditions, Advanced Composite Materials, https://doi.org/10.1080/09243046.2021.1918608*
20. Vijay Vasanth A,Latchoumi T.P, Balamurugan Karnan,Yookesh T.L (2020*) Improving the Energy Efficiency in MANET using Learning-based Routing, Revue d'Intelligence Artificielle, 34(3), pp 337-343.*
21. Wei-Nan Zhang, Zhao-Yan Ming, Yu Zhang, Ting Liu, and Tat-Seng Chua (2016) *"Capturing the Semantics of Key Phrases Using Multiple Languages for Question Retrieval", IEEE*
22. Yan-Hui Tu, Jun Du, and Chin-Hui Lee (2019) *"Speech Enhancement Based on Teacher-Student, Deep Learning Using Improved Speech, Presence Probability for Noise-Robust Speech Recognition",IEEE/ACM*
23. Yunshi Lan, Shuohang Wang, and Jing Jiang (2019) *"Knowledge Base Question Answering with a Matching-Aggregation Model and Question-specific Contextual Relations", IEEE/ACM*
24. Yuhua Lin, Member, Haiying Shen, (2015) *"SmartQ: A Question and Answer System for Supplying High-Quality and Trustworthy Answers",IEEE*