AUTOMATED QUESTION ANSWERING SYSTEM FOR EFFECTIVE E-LEARNING

Aathira Nair, Dipanjan Das, Prof. Deepti Lawand
Student, Student, Professor
Department of Information Technology
Pillai College of Engineering, New Panvel, India

Abstract: In today’s electronic age, the internet is widely being used for learning through E-learning websites which is totally a new and easy way of learning. E-learning is gaining importance day by day and emerging as a new way to deliver online education. In the current environment, tutors face a lot of difficulties and inconvenience in preparing answers which contain the detailed and correct information to each and every query or question that they receive from the students on a daily basis. This task takes up a lot of their time as the tutors have a very limited time frame to give suitable replies to all the questions asked each and every time. In order to solve this issue, we have developed a website which is a QAS i.e an Automated Question Answering System that helps to provide the most accurate answers to the questions of students automatically. In this, a system is developed in which the users who wish to learn and the teachers with the support of automatic agents integrate to find the most appropriate and meaningful answers to questions.

Index Terms - Question answering system, NLP, Information retrieval, e-Learning.

I. INTRODUCTION

From the past ten years, the percentage of internet users is increasing day by day. In Question Answering (QA) System, different queries are provided by the user with the aim of getting accurate answers. QA system provides a good solution to retrieve valid and correct answers to questions asked by internet users in the form of a query asked in natural language. Answers are formed by the Question Answer [QA] implementation (which is a computer program) by querying from a structured database of knowledge or information, which is a Knowledge Base. The QA system fetches answers from a natural language document collection which is basically unstructured in nature. Besides being a very specialized area in the field of information retrieval, this system is also convenient and time saving for the students who can get correct answers to their questions as well as beneficial to the tutors. We have also added a way to fetch answers from Wikipedia to those questions which are not present in the database. This makes the system more enhanced and easy to use.

II. LITERATURE SURVEY

In this literature survey, the relevant techniques are reviewed. It describes various techniques used in the following research paper work. Various techniques in this category are listed here. They are Neural Network Algorithms, Natural Language Processing, Information Retrieval, and Machine Learning. These techniques have various advantages and disadvantages which are listed below.


Some blocks may not contain the keywords of the query. However, they will be extracted as answer blocks till the time they contain the answer and have some similarity to other answer blocks. For an easy and quick comparison, the blocks having the answers are combined or grouped into a single page. The system can adapt to multiple resources as it is frequently available in the e-learning domain. Additional effort is needed to improve the speed and the system’s accuracy of prediction and to enable it to handle a very high workload.


This system is able to create systems that evaluate or grade the answers with results that are consistent with human performance. Since its application needs are high, QA system needs great potential for exploring challenges in this domain.

2.3 Bao J., Duan N., Zhou M., and Zhao T., “Knowledge-Based Question Answering as Machine Translation”, 2014

This system integrates semantic parsing through a translation based and knowledge-based question answering system in a single framework. Using an independent semantic parser with state of the art performance, it is observed that the system achieves better results on a general domain evaluation set.


The same knowledge is learned by the System from data. The difference is that in the IE approach the knowledge is acquired through a direct connection between dependency parses and answer properties, while in the case of SP approach, it is acquired through optimizing intermediate logic forms. There is a need to create more compositional open domain datasets. SP researches have to concentrate on utterances in existing or present datasets that are out of reach of direct IE methods.

It is a Real-time and Offline system. Real-time for this system means the system tries to find an answer from the online user. The user receives a message in the form of a popup dialogue. Students can improve their skills and correct problems together with their friends without waiting for help from the instructor. The students from various levels of ability are able to study the lesson on their own. The system can work only with a particular university’s agent. It is not available for everyone.

III. EXISTING SYSTEM

The existing system uses a Support Vector Machine. SVM is a linear discriminate model that finds a hyper plane with a maximum margin for separating the classes. They are fast classifiers to classify questions for high dimensional data which are trained by around 1000 labeled questions. The question classification SVM achieves above 80% accuracy. It is an open domain system. The main aim of this system is to permit the computers to learn automatically without human intervention or assistance and adjust actions accordingly. The source of knowledge for this Question Answering System is through the web. That is all the answers generated in the system come from Wikipedia with several options of answers. The system architecture is separated into two modules which are Question Processing Module and Answer Extraction Module. The Question Processing Module consists of Question Classification, Question Focus Identification, and Query Expansion. Answer Extraction Module consists of Document Retrieval, Passage Retrieval, Answer Selection and Presentation and Answer Corpus.

IV. PROPOSED SYSTEM ARCHITECTURE

In this proposed system, there is an admin panel which deals with uploading the videos/documents. The admin takes note of maintaining the system and website and has access to all the confidential details about the website. Users can log in to the website and can view the videos when needed. After viewing the video, if the user has any doubts regarding the video, they can get it clarified by asking that particular question and the answering system generates answers from the database which is stored with relevant questions and answers. In this system, Seq2Seq model is being used.

4.1 Database:
In this system’s database, all the data which are important like contact us, user login information, and student details get saved in MySQL of the database. All the data information would be driven from the MySQL database. Datasets are stored and saved in the JSON file.

4.2 Server:
The server used in this system is the Python server. Coding is done in python. We have used Jupyter notebook which is an open-source web application. It shares and creates documents that contain equations, live code, and visualizations. Through this, we will get the output to the questions asked directly then and there. We can also run each code individually with Jupyter. The code is saved in the “.ipynb” extension.

4.3 Admin:
The Admin has to upload audios/videos/documents from time to time to enable the users to view it and make maximum use of it for learning. The duty of the Admin is to maintain the website, its configuration and ensure reliable operation of the computer systems, especially servers which are multi-user computers. The Admin has to ensure that the requirements of the users are fulfilled and they can benefit from the website.

Fig. 1. Proposed System
4.4 Answering System:
In this E-Learning System, using the fields of Deep Neural Networks, Machine Learning, Information Retrieval and Natural Language Processing [NLP], the user will be able to automatically get relevant and accurate answers to questions asked by them in a defined language. The answer to the question posted will be accurate and meaningful instead of various options as available in Wikipedia where each answer has a different meaning.
However, in case there is no question to the answer asked, then it will fetch those answers from Wikipedia. So the system becomes more unique and developed because of both the options.

4.5 User Interface:
It shows the website’s view as a whole. It refers to the appearance of the website to the user. The main objective of a Graphical User Interface [GUI] is to help users to navigate through the website in an easy manner with minimum trouble. While using the Website, it indicates to the user where they are at present in the course which means how much of the course they have completed and what is remaining to be completed. The GUI of this website has keys such as Play, Replay, Previous, Pause, Next, Glossary, Resources, Audio On / Off, Exit, etc. which helps the learner to move easily through the course content. The user can easily interact with the UI which is in the form of an Information Device. The UI includes keyboards, mouse, display screens, and the desktop appearance which enables the user to meaningfully interact with an application or website.

V. IMPLEMENTATION DETAILS

5.1 Text pre-processing:
The first step in the implementation of the question answering system is text pre-processing where the full text is converted into lowercase or uppercase so that the algorithm does not fetch the same words in different cases as different words.

5.2 Tokenization:
The next step is tokenization where normal text strings are converted into a list of tokens. To locate the list of sentences, sentence tokenizer is used and for a list of words, word tokenizer is used. Stream of words is broken into words, symbols, phrases and other elements to form tokens. NLTK data package already includes a pre-trained Punkt tokenizer for English.
Example: from nltk.tokenize import sent_tokenize
text = "Hello everyone. Welcome to SmartHub. You are studying nlp article"
sent_tokenize(text)
Output: ['Hello everyone.', 'Welcome to SmartHub.', 'You are studying nlp article']

5.3 Removing noise:
All the words or texts that are not in a standard number or letter format are removed from the sentence which is known as removing noise. The very common words which appear to have very less value in helping to select the documents matching the need of the user are removed and neglected from the vocabulary which is known as Stop words.

5.4 Stemming:
In stemming, all the suffixes are removed from the word and are reduced to the root node. In this process, derived or inflected words are reduced making it a single word as the machine does not understand English grammar and will not differentiate the verb. For example: For the word “Waiting”, its suffix is “ing” and if we remove “ing” from “Waiting”, we will get the base word or root word which is “Wait”.

5.4 Lemmatization:
Lemmatization is an algorithmic process in which suffixes are added to the words. Lemmas are actual words. So the word that you end up which is the root stem are not something that you can search in a dictionary but you can look up a Lemma to locate the word. For example: “wait” is the base form for words like “waiting” or “waits” or the word ”better” and ”good” are in the same lemma, hence these words are considered the same.

5.5 Keyword matching:
Keyword matching is the same as the concept used by ELIZA which is a simple keyword matching for greetings. In this process, the answering system returns basic greeting responses if the user's input is a greeting.

5.6 Bag of words:
The words in the database are the representation of texts that tells you the occurrence of words in the document which are known as Bag of Words. It consists of a vocabulary of known words. This model focuses on the known words that occur in the document and not where they occur in the document. The meaning of the document is understood by the content alone. For example: {Life, is, the, not, great} is present in the dictionary and we want to vectorize the text “Life is great”, then we would have the vector as: (1,1,0,0,1).

5.7 Term frequency/ TF-IDF approach:
By the TF-IDF approach which is also known as Inverse Document Frequency, the frequency of the words is rescaled by how many times they appear in the documents so that the scores for the frequent words like “the” that are also frequent across all the other documents are penalized. The score of the frequency of the words in a particular document is termed as “Term Frequency”.
TF = (Number of times term t appears in a document)/(Number of terms in the document)
Scoring of how rare the word is across the documents is done by Inverse Document Frequency.
IDF = 1+log(N/n), where, N is the number of documents and “n” is the number of documents a term t has appeared in.
Cosine similarity: Tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

Cosine Similarity \( (d_1, d_2) = \text{Dot product}(d_1, d_2) / \|d_1\| * \|d_2\| \) where \( d_1, d_2 \) are two non zero vectors.

**VI. INPUT & OUTPUT DETAILS**

The input for this system is in the form of words and sentences. That is, whenever you come across any doubts regarding the video, you can ask by just typing your question. The answering system fetches the possible answers by getting its probability and matching it with the database.

Ex.: Types of digital forensics?
- What is cryptography.
- Definition of Digital forensics.
- What is cyber stalking?
- What is nlp.

The output shows a set of questions and answers which are the queries or doubts of the users viewing the video or document. The reply of their doubts is given automatically by the highest probability order from the given set of questions and answers in the database stored. For example when we type a doubt as “What is cybercrime”. Then the answering system would reply it as “Any illegal activity committed using a computer and/or the internet can be called a cybercrime”.

**VII. RESULTS/SCREENSHOTS**

![Smart Hub website](image1)

Fig. 2. Smart Hub website

The above image shows the working of the question answering system wherein when you view the video and ask questions related to the video, the automated answer gets generated.

![Question answering system](image2)

Fig. 3. Question answering system

These automated answers which are generated come from the datasets which are stored in the JSON file.
VII. CONCLUSION

The Automated Question Answering system with effective E-learning using Deep Neural Networks, NLP, Machine Learning, and Information Retrieval is developed to enable the user to extract the most accurate and appropriate answers to their queries on the basis of the videos/texts uploaded in the system where the answering system in the server automatically generates answers for the queries instead of depending on actual tutors for getting answers which are quite time-consuming. The system can be designed as per individual module and they could easily reside on the servers. It is very beneficial to students who can't afford to go to tutoring or coaching classes.

IX. ACKNOWLEDGMENT

We would like to thank our mentor, Prof. Deepti Lawand for her guidance and unwavering support throughout the project and the semester. We would also like to thank our HOD, Dr. Satishkumar L. Varma for his encouragement and motivation to learn and implement projects of sorts. Lastly, we would like to thank our Principal, Dr. Sandeep Joshi for providing us opportunities to explore our domain and for motivating us to do better.
REFERENCES


