Smart Education Chatbot – A Chatbot Based Question Answering System for Retrieving Answers From PDFs Using BERT

Prakash Sewani
Department of Computer *Engineering*
WIEECT Ulhasnagar (West),
Mumbai, India,
prakashsewani1994@gmail.com

Bhavesh Thadhani
Department of Computer *Engineering*
WIEECT Ulhasnagar (West),
Mumbai, India
bhavi.thadhani@gmail.com

Sandeep More (Mentor)
Department of Computer *Engineering*
WIEECT Ulhasnagar (West),
Mumbai, India
sandeepmore@mail.watumull.edu

Arman Budhrani
Department of Computer Engineering
WIEECT Ulhasnagar (West),
Mumbai, India,
armanbudhrani007@gmail.com

Abstract— The use of Chatbots has evolved rapidly in numerous fields in recent years, including Marketing, Supporting Systems, Education, Health Care, Cultural Heritage, and Entertainment. Higher education comprises an important field for the application of chatbots, especially for large-scale use. In the proposed system, we present a new way of answering queries (questions) asked by users. The proposed system identifies the user context which triggers the particular intent for a response. Since it is responding dynamically, a desired answer will be fetched for the user. The proposed system utilizes a popular Question Answering algorithm by Google known as BERT. It is a context-based question answering algorithm that tokenizes given queries and generates using modern NLP techniques. The proposed system also utilizes the concepts of context identification and web scraping.

Keywords— Chatbots, Context Identification, Web Scraping, BERT, NLP.

I. INTRODUCTION

A computer program, algorithm or artificial intelligence which communicates with a person or another participant of communication can be called a 'chatbot'. Chatbots comprise computer programs that are used to simulate auditory and/or textual conversations with users, or chatbots using natural languages. Chatbots – also machine conversation, virtual agent, dialogue system, and chatterbot – commonly appear in customer services responding to frequently asked questions (FAQs) and offering technical support, in business webpages for selling products, and as personal assistants on mobile devices. They are becoming a trend in many fields such as medicine, produce, and service industry and lately in educations.

Chatbots' conceptualization emerged from the need of human to interact with computes in a natural human language. A key milestone in the pre-history of chatbots holds in the so-called Turing Test. In the 1950s, the possibility of training a computer machine to launch conversations with users led Alan Turing proposing the Turing Test; a program that developed a text message conversation with a prober for five minutes.

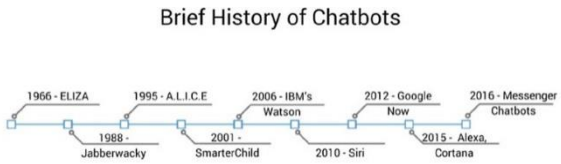


Fig.1 History of Chatbots

Following this, ELIZA, the first chatbot in the history of Computer Science was created in 1964-1966 by Joseph Weizenbaum at Massachusetts Institute of Technology (MIT). ELIZA used simple pattern matching and a template-based response mechanism to imitate the conversational style of a nondirectional psychotherapist, in the early scenario called DOCTOR. The use of chatbots increased dramatically with the massive expansion of the Internet and especially social networking sites, Chatbots have been further advanced in the last decade, due to development in natural language processing and machine learning algorithms, such as deep learning and neural networks which perform Artificial Intelligence (AI) tasks like Image Recognition, Natural Language Generation, Speech Recognition and Text to Speech Synthesis.

Chatbots in Education

In recent decades, the number of students per lecturer has constantly risen. Large-scale lectures at universities with more than 100 students per lecturer and massive open online courses are increasingly becoming the default learning scenario. Consequently, individualized support provided by lecturers is nearly impossible and students are unable to engage in effective learning. Several studies have revealed that this lack of individualized support leads to weak learning outcomes, high dropout rates and dissatisfaction. The best solution would be to have one teacher per student. Obviously not possible due to financial and organizational restrictions.

Chatbots have the potential to solve this problem using the examples of other sectors. Chatbots have a growing presence in modern society, becoming integral parts of everything from personal assistants on mobile devices to technical support help over telephone lines, and even being used for health interventions.

While the presence of chatbots, on web platforms and/or standalone applications is already substantial in customer services, business webpages, product sales, and in health interventions, their use in education is still in its infancy. A meaningful integration of chatbots in higher education presupposes a good understanding of user's needs and expectations, as well as, an examination of their perceptions towards education technology, the adoption of an appropriate pedagogy and last but not least, a confrontation of technological challenges and potential limitations, that relate to the Natural Language Processing (NLP) research field.

II. LITERATURE SURVEY

A. *Algorithms:*

- **YAKE!**

YAKE! (Yet Another Keyword Extractor) is a light-weight unsupervised automatic keyword extraction method which rests on text statistical features extracted from single documents to select the most important keywords of a text. The system does not need to be trained on a particular set of documents, neither it depends on dictionaries, external-corpus, size of the text, language or domain.

```
import yake

text="what is newton's first law?"

kw_extractor = yake.KeywordExtractor()
keywords = kw_extractor.extract_keywords(text)

for kw in keywords:
    print(kw)
```

Fig.2 YAKE! Algorithm

```
PS C:\Users\praka\Desktop\Watumull> & C:/Users/praka/AppData/Local/Programs/Python/Python39/python.exe c:/Users/praka/Desktop/Watumull/nlpt.py
('newton first law', 0.04940384002065631)
('law', 0.15831692877998726)
('newton', 0.29736558256021506)
PS C:\Users\praka\Desktop\Watumull>
```

Fig.3 YAKE! Output

- **BERT**

BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. In NLP, this process is called attention.

BERT model is designed in such a way that it can read text from both directions, i.e., from left-to-right and right-to-left. Using this bidirectional capability, BERT is pre-trained on two different but related NLP tasks: Masked Language Modelling and Next Sentence Prediction

The objective of Masked Language Model (MLM) training is to hide a word in a sentence and then have the program predict what word has been hidden (masked) based on the hidden word's context. The objective of Next Sentence Prediction training is to have the program predict whether two given sentences have a logical, sequential connection or whether their relationship is simply random.

Next Sentence Prediction:

In the BERT training process, the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document. During training, 50% of the inputs are a pair in which the second sentence is the subsequent sentence in the original document, while in the other 50% a random sentence from the corpus is chosen as the second sentence. The assumption is that the random sentence will be disconnected from the first sentence.

To help the model distinguish between the two sentences in training, the input is processed in the following way before entering the model:

1. A [CLS] token is inserted at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence.
2. A sentence embedding indicating Sentence A or Sentence B is added to each token. Sentence embeddings are similar in concept to token embeddings with a vocabulary of 2.
3. A positional embedding is added to each token to indicate its position in the sequence. The concept and implementation of positional embedding are presented in the Transformer paper.

To predict if the second sentence is indeed connected to the first, the following steps are performed:

1. The entire input sequence goes through the Transformer model.
2. The output of the [CLS] token is transformed into a 2×1 shaped vector, using a simple classification layer (learned matrices of weights and biases).
3. Calculating the probability of IsNextSequence with softmax.

B. *Literature Review:*

The state of the research on chatbot applications in the education sector was described with according to the reviewed articles. A review of the studies selected for this survey found that chatbots were used in a variety of ways, including teaching and learning, administration, evaluation, advice and research and development, for educational purposes. The introduction of learning pedagogy as the chatbot system in education has personalized online learning and made student learning materials accessible anywhere and anytime. According to their studies, chatbots are good technological innovations that can improve student engagement, cognitive acquisition, and performance.

Today's chatbots typically use two types of models to generate responses, a retrieval-based model and a generation-based model. An on-demand model is a common

choice of chatbot because it is easier to develop and maintain, e.g. For example, this model is essentially a matching process between the input question and an output answer. We can use either cosine similarity or a trained CNN, LSTM model to calculate the probability of a match between a question and an answer. All responses from bots are maintained and stored in databases. When an input question is received, this model calculates the probability of a match for the input question and for all responses in the database. The higher the probability of a match, the more likely the correct answer is. Google's email autoresponder suggestion system is based on, the same type of on-demand model. This model is ideal for a quality control system that needs precise and exact answers, because the answers are stored in databases. The quality of responses can also be guaranteed. However, this Model suffers from poor response coverage because natural language is complex. You cannot answer a question or enter a phrase that is not included in the predefined answers. Creating a database containing all kinds of topics, responses and answers is time consuming, difficult to maintain, and difficult to cover all possible responses in natural language. Unlike a retrieval-based model, a generation-based model does not store any predefined responses in a database. Generates a response through the model itself so that it can be used in open or wide coverage applications. The most commonly used generation-based machine learning model is sequence-by-sequence models. Sequence-by-sequence models, known as seq2seq, consist of an encoder and a decoder based on a neural network model. In the training state, the encoder receives word embedding vectors from the question mark string as its input, and decoder receives the word embedding vectors from the response string as its output. By coupling the encoder vector output to the input of the decoder, we can train the model to generate responses through its trained decoder. The retrieval-based model focuses on improving the learning efficiency of E-Learning and the other everyday tasks of a personal assistant.

One of the approaches to this project is [1] paper, in which Farhan M. et al. using a web bot in an e-learning platform, to address the lack of real-time responses for the students. In fact, when a student asks a question on e-learning platform the teacher could answer at a later stage. If there are more students and more questions, this delay increases. Web bot is a web-based Chatbot that predicts future events based on keywords entered on the Internet. In this work Pandora is used, a bot that stores the questions and answers it on XML style language i.e., Artificial Intelligence Markup Language (AIML). This bot is trained with a series of questions and answers: when it cannot provide a response to a question, a human user is responsible for responding. In the last recent years some interesting research works can be found. This was an interesting approach to an Educational Chatbot as it utilised the essentials of Machine as well as Human together.

In [2] paper, Satu S., Chatbot is called Tutorbot because it is functionality backing of didactics done in eLearning environments. It contains some features as natural language management, presentation of contents, and interaction with search engine. Besides, e-learning platforms work is linked to indispensable services to web service. This project utilises various Educational Website APIs, search engines and backend data to provide an appropriate answer to the students.

In [3] paper, Muhammad Rana, while developing EagleBot, categorized user queries into three parts as Unstructured QA, Structured QA, FAQs using Google's Dialogflow. This enabled him to easily fetch answers for the user. For Structured QAs he usually fetched the queries from the data stored in databases. For Unstructured QAs, Muhammad Rana used the information present with the regarding query and web scrapped some results as well and fed it directly to the BERT Algorithm along with user query. This returned the appropriate result. For FAQ QAs, he web-scrapped his University Website for similar topics of discussions and provided user with the links to the forums.

Francesco Colace, Massimo De Santo, Marco Lombardi, in their paper "Chatbot for E-Learning: A Case of Study", presented Latent Dirichlet Allocation Algorithm and Workflow Manager to provide information about the courses offered by the said University using their Knowledge Base consisting of articles, user queries and answers based on the University Website.

III. LIMITATIONS

LIMITATIONS OF PYPDF2 MODULE:

PDF reading algorithms utilizes PDFs that already contain text in them. They cannot read images. Another drawback of using such algorithms is that when faced with an equation in the PDF, if proper encoding is not used, the equation will not be represented properly. If encoding is used to solve this error, text errors might occur while dealing with special characters such as " ' ", " ^ "and so on.

LIMITATION OF BERT ALGORITHM:

BERT algorithm utilizes context for answering user queries. The context provided to the algorithm are faced with certain rules. BERT uses word-piece tokenization. So, when some of the words are not in the vocabulary, it splits the words to its word pieces. For example: if the word playing is not in the vocabulary, it can split down to play, ##ing. This increases the number of tokens in a given sentence after tokenization. Max number of tokens that is algorithm can handle is 512. This is more than enough for question answering but sometimes the context itself is not enough for the answer which leads to improper responses.

IV. PREVIOUS WORK

In one study, an android application was presented containing education chat-bot designed for visually impaired people (Kumar et al., 2016). The application can be launched with Google Voice Search and is used by asking questions in spoken natural language. The application then converts it into text and run the query against the AIML database or, if no pre-saved answer was found, against the Wikipedia API. Though the application of this chatbot is noble, the retrieval architecture is a simple rule-based one.

Jill Watson, the Georgia Tech teaching assistant chatbot demonstrated the strong viability of chatbots in the

educational domain (Goel and Polepeddi, 2016). Jill Watson showed promise as an alternative to teachers in the near future. Georgia Tech Computer Science Professor Dr. Ashok Goel built this chatbot to help students on their assignment related questions in one of his Artificial Intelligence courses. JW1 (Jill Watson version 1) was built using IBM Watson APIs. JW1 had a memory of question answer pairs from previous semesters organized into categories of questions.

Aforementioned chatbot systems work for a very specific domain and don't deal with the whole university domain. As of now, as to the best of our knowledge, there is no such integrated system for answering all types of questions asked in the university domain.

V. IMPLEMENTATION

TECHNOLOGY

Technologies used:

- Flask
- HTML/CSS
- JavaScript
- BERT
- Web Scraping

PROPOSED METHODOLOGY

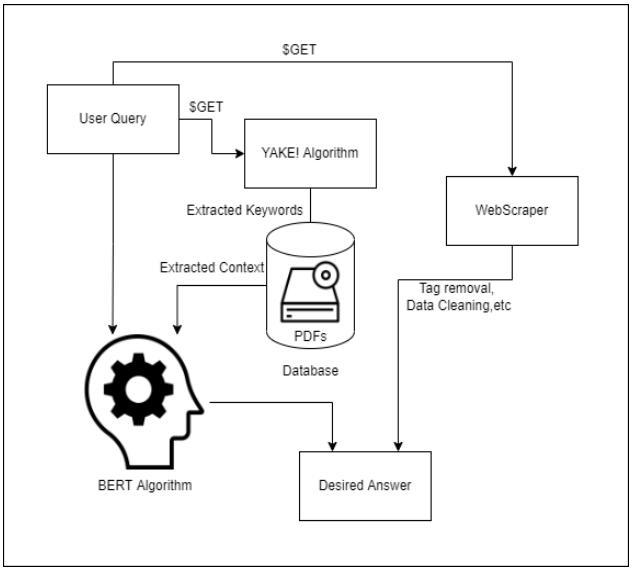


Fig.4 Blueprint of Proposed System

TECHNOLOGY:

Technologies used:

- Flask
- HTML/CSS
- JavaScript
- YAKE!
- BERT
- Web Scrapping
- Wikipedia

Current Configuration:

- CPU: AMD Ryzen 5 3450H
- GPU: Nvidia GTX 1650Ti
- RAM: 8Gb
- Storage: 256Gb

Recommended System:

- CPU: AMD 3rd Gen Duo Core or Intel Equivalent
- GPU: Not Required
- RAM: Minimum 3 Gb Free
- Internet Connection: Not Mandatory

IMPLEMENTATION

User queries are fetched using the GET Method in Python Flask. These queries are then passed to the YAKE! Keyword Extraction Algorithm to extract the necessary keywords into a variable.

These keywords are then passed through the PDFParser Algorithm which extracts the necessary context for the BERT Algorithm

Fig.5 Context Extraction Algorithm

This context along with the User Query is passed on to the BERT Question Answering Algorithm which tokenizes the context into words using its own tokenizer. Let's the word "playing" is not in the vocabulary, it can split down to play, ##ing. This increases the number of tokens in a given sentence after tokenization. Max number of tokens that is algorithm can handle is 512. This is more than enough for question answering but sometimes the context itself is not enough for the answer which leads to improper responses.

BERT Algorithm then returns the answer by relating the tokens and the User Query using Contextualized Embedding, meaning, making sense of the two sentences

VI. OUTPUT

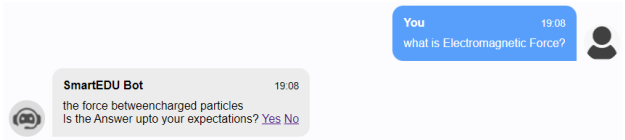


Fig.6 BERT Output

Pressing the "No" button leads to the user query being fitted into the WebScraping Module. This module scrapes the top website provided by Google Search and returns the answer that fits the best

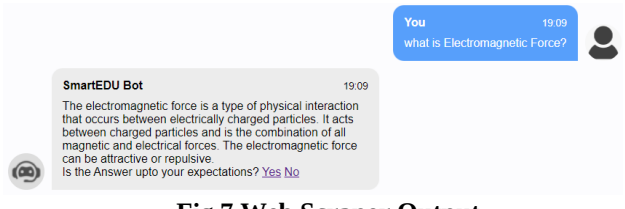


Fig.7 Web Scraper Output

VII. TESTING AND RESULTS

Test Cases and Accuracy of our Chatbot are stated below:

User Asked: "What is Gravitational Force?"
Expected Answer: "The gravitational force is the force of mutual

attraction between any two objects by virtue of their masses"
Calculated Answer: "the force of mutualattraction between any two objects by virtue of their masses."

User Asked: "What is Newton's Third Law?"
Expected Answer: "To every action, there is always an equal and opposite reaction."
Calculated Answer: "newton ' s third law ?"
Web scrapped Answer: "Newton's third law states that when two bodies interact, they apply forces to one another that are equal in magnitude and opposite in direction. The third law is also known as the law of action and reaction."

User Asked: "What classes come under Phylum?"
Expected Answer: "Classes comprising animals like fishes, amphibians, reptiles, birds along with mammals constitute the next higher category called Phylum."
Calculated Answer: "fishes , amphibians , reptiles , birds"

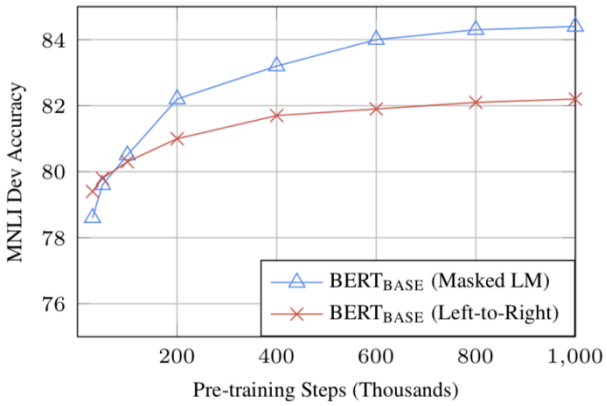
User Asked: "What is Genus?"
Expected Answer: "Genus comprises a group of related species which has more characters in common in comparison to species of other genera."
Calculated Answer: "a group of related species"

User Asked: " "
Expected Answer: " "
Calculated Answer: "Error- List Index Out of Range"

As you can see the answers provided by the BERT algorithm may not be according to the University standards but they are to the point, without mention of phrases like "This term Means", "This term includes". All the answers comprise of words that co-relate to the user query and are to the point.

Accuracy of BERT Algorithm:
The max number of intents that our Fine-Tuned BERT model can handle is 512. This number can be lower, can be equal to 512 but not greater than it since the model is not able to handle more than 512 intents. BERT Model uses its own Tokenizer, this means it has its own vocabulary. Let's say you pass the word "playing" as context for BERT Algorithm. The vocabulary of BERT Algorithm contains the word "play" not the word "playing". It will hence split the word "playing" as "play" and "#ing". This adds up to 2 intents. Context should be simple enough to be understood and not to complex with words outside the vocabulary.

The calculated accuracy for the BERT Algorithm with 381 intents is 86.11%. This accuracy keeps on decreasing by a logarithmic curve as the number of intents increase. There is no fix accuracy for BERT Algorithm since intents will be different every time and two sentences can always never have the same meaning.



VIII. CONCLUSION

From the research conducted above we aim towards creating a perfect Smart Student Assistant. We are using advanced data parsing libraries such as PyPDF2 to extract information from various PDFs and send it to our backend. We are also using Wikipedia API and some Educational APIs that shall also provide data to our backend. By using this Bot, we assure that the student will be able to find both short and brief answers for the asked question.

Future Scope:
We aim to implement image processing so that images can be exchanged by the bot as well as the user. This will enable them to get appropriate representation for their problem.

We also aim to implement a dynamic PDF upload option so that you, the user can upload your own PDF and start asking questions related to the topics present in the PDF right away.

IX. ACKNOWLEDGEMENT

We thank Prof. Sandeep More for helping us in creating this research paper and guiding us throughout this project.

X. REFERENCES

[1] Rana, Muhammad, "EagleBot: A chatbot based on Multi-Tier Question Answering Systems for Retrieving Answers from Heterogenous Sources Using BERT" (2019). Electronic Theses and Dissertations.

[2] Francesco Colace, Massimo De Santo, Marco Lombardi, Francesco Pascale, and Antonio Piertrosanto, "Chatbot for E-learning: A Case of Study", DIIN University of Salerno, Fisciano (SA), Italy

[3] K. Jwala G.N.V.G Sirisha, G.V. Padma Raju, "Developing a Chatbot using Machine Learning", Electronics Computer Technology (ICECT) 2011 3rd International Conference on, Vol. 4, 2011

[4] Ashok K. Goel and Lalith Polepeddi, "Jill Watson: A Virtual Teaching Assistant for Online Education", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-IS3, June 2019.

[5] Sam Schwager John Solitario, "Question and Answering on SQuAd 2.0: BERT is All You Need" 2016 15th International Conference on Department of Computer Science Stanford University

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Farhan M. et al. "Automated Web-Bot Implementation using Machine Learning". Journal of Applied Environmental and Biological Sciences.