**Paragraph Question and Answering Application NLP**

***Abstract:*** A question and answering (Q&A) application that uses natural language processing (NLP) is a system designed to interact with users in a natural, human-like way to provide accurate and relevant answers to their queries. NLP is a field of computer science that focuses on the interaction between computers and humans using natural language, and it has been increasingly applied in the development of chatbots, virtual assistants, and other conversational agents. The Q&A application using NLP works by first identifying the user's intent behind the query, then extracting relevant information from a variety of sources such as databases, websites, and documents, and finally generating a response that is both accurate and understandable to the user. This process involves several stages, including natural language understanding, information retrieval, and natural language generation. Overall, a Q&A application using NLP has the potential to provide a seamless and efficient way for users to get answers to their queries. By leveraging the power of NLP, such an application can provide accurate and relevant answers to a wide range of questions, helping users to find the information they need quickly and easily.

***Keywords:*** Question and answering, Natural language understanding, Natural language generation, Information retrieval.

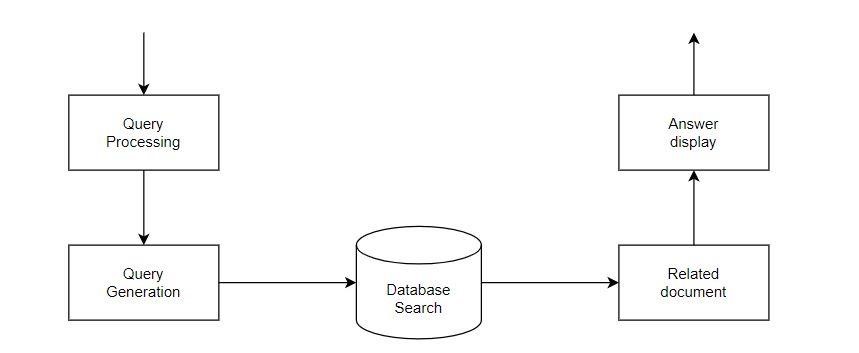
1. ***Introduction:***

In this project, we will be developing a Q&A system using NLP that can understand and respond to user queries accurately and efficiently. Our system will be designed to process natural language input, identify the user's intent, and retrieve the most relevant answers from a given dataset. We will be using a combination of NLP techniques, including text preprocessing, entity

recognition, intent classification, and answer generation, to build an intelligent Q&A system.

Natural language understanding (NLU) is the ability of the system to comprehend the meaning of the user's query. This involves breaking down the query into its constituent parts, such as identifying the key nouns, verbs, and adjectives, as well as any relationships or dependencies between them. NLU also involves identifying any entities mentioned in the query, such as people, places, or things, and determining the context in which they are being mentioned. Information retrieval (IR) is the process of finding relevant information to answer the user's query. This involves searching through various sources of information, such as databases, websites, and documents, to find the most relevant and accurate answers to the user's question. IR can be done using various techniques, such as keyword matching, semantic analysis, and machine learning.

Natural language generation (NLG) is the process of generating a response that is both accurate and understandable to the user. This involves taking the relevant information retrieved through IR and generating a response that is grammatically correct, logically coherent, and tailored to the user's level of understanding. NLG can be done using various techniques, such as template-based generation, rule-based generation, and machine learning. One such technology that has gained popularity in recent years is natural language processing (NLP), which is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and human language. NLP has been applied in various domains, such as machine translation, speech recognition, sentiment analysis, and question answering systems. In this context, a question and answering (Q&A) application that uses NLP can provide a seamless and efficient way for users to get answers to their queries in a natural, human-like way. This paper will provide an overview of Q&A applications that use NLP, including their architecture, working principles, and potential applications.



**Figure1. Architecture of question-answering system**

1. ***Literature Review:***

The author of article [1] created a geographical domain question answering system that provides answers to user questions about various cities. To develop the system, the first author generates a knowledge base document and uses Named entity Tagger, Parser, and Word Net tool to execute document pre-processing such as noise removal, tokenization, sentence splitting, and document tagging. The major components of this system are question processing, document processing, and response processing. Question processing is concerned with question sub-classification and reformulation. A simple matching pattern technique is used to classify questions. Following that, there was passage retrieval, which employed a pre-processed and indexed corpus for passage retrieval. The retrieval module generated candidate answers, which were then utilized as input for answer extraction, with ranking based on semantic relationship using the Word net tool. After rating, the final solution is displayed with the highest rank. The author presented IPedagogy in article [2], which is a question answering system that works with natural language driven queries to get responses from specified information clusters by minimizing the search area of information retrieval. Furthermore, IPedagogy is supported by a number of natural language processing algorithms that lead the system to retrieve the exact response to a given inquiry. The system is tested using mean reciprocal rank, and it is discovered that the system has an average accuracy level of 0.73 for 10 sets of 35 questions. The author of article [3] created a novel architecture for a Malayalam question answering system that discovers solutions to Malayalam queries by analyzing Malayalam documents. It deals with the four types of Malayalam questions for closed domain. Question Analysis, Text Retrieval and Answer Snippet Extraction, and Answer Identification are the three components of the system.

1. ***Methodology:***
2. **Question Analysis:**

It only accepts specific searches as input. The purpose of this module is to detect questionrelevant keywords and predict the responses that go with them. The pre-processing procedure perfectly follows the NLP algorithm.

1. **Text Retrieval and Answer Snippet Extraction:**

Once the query phrases have been detected, the answer candidates are taken from the collected document for answer identification. For the answer extraction procedure, indexed papers with a complete keyword match to the question's keyword are picked. This is accomplished by counting the number of times the query term appears in each sentence. To represent the sentences that have at least a partial match to the query keyword as response possibilities, a triplet consisting of the sentence, index, and count of the match is utilized. The index is used to retrieve the actual sentence. The index term of text splitting and match count is now assigned to the phrase that matches the query. These replies are presented to the following applicant.

1. **Answer Identification:**

Answer extraction, as well as scoring and ranking, are two of the answer identification sub-modules. The winner is determined by matching window sizes while scoring and rating the participants' responses. The candidate response with the highest score is selected as the winner candidate, and the results of the winner candidate are processed to extract the answer. The expected named entity is determined by analyzing the question word, and then the question word that is closest to the surrounding word is analyzed to establish the expected response entity.

1. **Question Processing Module:**

The user Question is handled in the Question Processing module in order to extract some vital information from it. The steps by which the query Processing Module is processed are listed below. Question Processing Module Procedures:

* 1. Determine the type of inquiry by using the Wh word.
  2. Determine the expected type of response.
  3. Extract the question's keywords. **E. Document Processing Module:**

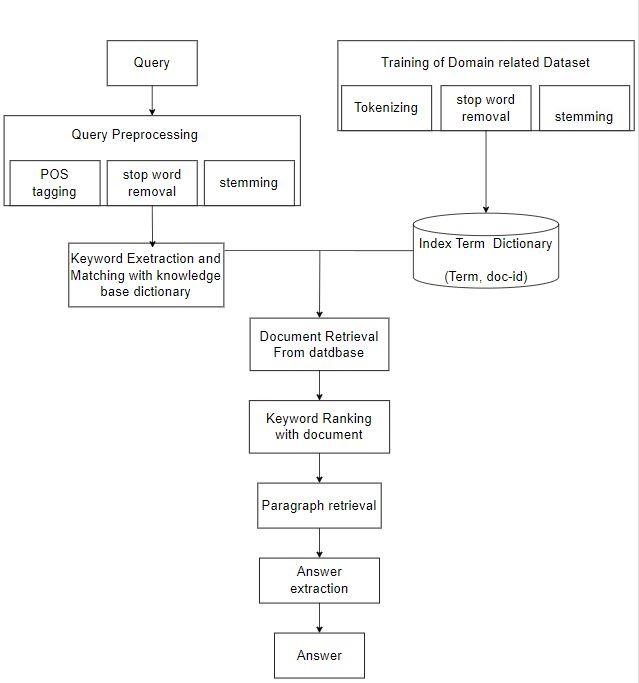
The papers relevant to the specified inquiry are obtained and processed in the document processing module. Steps in the Document Processing Module:

* 1. After processing the query, use a trusted search engine to find related documents to that topic.
  2. Select the 10 most relevant documents.
  3. Extract the required material from these papers.
  4. Save the information in a text file. **F. Paragraph Extraction Module:**
  5. This is the module where the tasks of Paragraph Extraction and Sentence Extraction are carried out in order to determine the most likely answer to the topic at hand. The following are the steps in the Paragraph Extraction Module:
  6. Run the Document Retrieval module's paragraph extractor on the text file.
  7. If the query is about a definition or a factoid, pass the extracted text to the next submodule.
  8. If the query is about a definition or a factoid, forward the extracted text to the next submodule.

**G. Answer Extraction Module:**

The algorithms discussed in this subject address three categories of questions: definitional questions, descriptive questions, and factoid questions. The author identifies the grammatical structure of the query using the Stanford Parser Toolkit and, based on that structure, determines prospective solutions from the dataset, which are sentences with the same head words as the inquiry.

1. ***Proposed Approach:***



**Figure2. Proposed architecture of QA**

The system will receive input from a query on educational legislation or other educational-related data. "What is the parent's responsibility to ensure their children's education?" as well as "What are the school funding authorities?" The Question keyword is computed by removing stop words and stemming on the question to obtain the answer. Using the collection of papers related to the Education Act, a metadata knowledge base with corresponding keywords for each document is constructed. The original text or phrases are marked with these keywords in order for the answer extractor to generate probable replies. According to the instructions, the constraint and candidate solutions are evaluated, and the most likely answer with the highest score is chosen as the correct response. The system will provide an accurate response to questions that have been taught before testing to determine how accurate untrained queries are.

**V*.Details of Implementation Methods:***

**1. Study of the Data Set**

The collection and analysis of relevant data is part of the proposed work's initial design module. The main data set for the endeavor is the records of education actions associated data. The essential data set is obtained and reviewed from various educational legislation sources. The system's next module, which is observation-based, generates a corpus. **A. Preprocessing**

Following the creation of the corpus, several preprocessing procedures are conducted on each corpus text file. Stop words and stemming are important responsibilities in preprocessing. **a. Stop Words Removal**

The removal of stop words reduces the dimensionality of term space. Prepositions, articles, and pronouns are the most regularly seen terms in text documents, although they do not give the meaning of the writings. Stop words are removed from publications since they are not considered keywords in information retrieval systems. For example, the English stop words "is, for, the, in, etc." are removed from each text file in the dataset by maintaining an English stop word dictionaries.

Stemming

1. **Index Term Dictionary:**

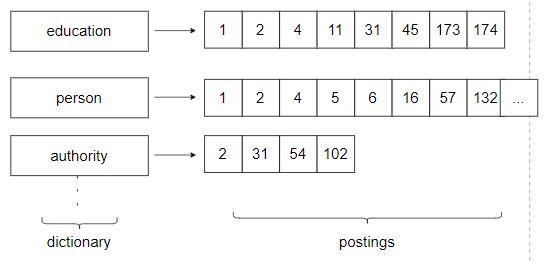
The extracted keywords are kept in the index term dictionary after preprocessing. Only those words that are obtained by stemming are included in the extracted keywords. Java is used to generate the index word dictionary, which is then saved in MySQL as a table. The structure of a dictionary is like an invert index, with two columns for terms and postings. The only difference between term and posting is an extracted keyword. The structure of the index word dictionary is depicted in the following image.

**Figure3.** **Structure of index term dictionary**

1. **Question Preprocessing:**

The supplied input query is being pre-processed by executing several preprocessing operations on it, such as POS tagging, stop word removal, and stemming.

**a. User Query**



The user will enter a question on the educational system. For example, the user may inquire, "What is the parent's responsibility to ensure their children's education?" or which elementary school?

Or any question about the educational system. **b. Pos Tagging**

First, we conduct POS (part of speech) tagging on the input query, labelling each word of the user question with its type, such as verb, noun, and so on. The POS stand ford tagger is used to tag each word.

**c. Extracted Keyword**

The keywords are collected from the user query. These keywords are obtained by eliminating the symbols and stop words from the user query; moreover, stemming is applied to keywords in order for them to match with index terms and dictionary terms for document retrieval. To extract keywords, an English stop words and stemmed words dictionary is kept. As an example:

**4.Document retrieval:**

The extracted keywords returned by a query in a document retrieval system correspond to terms in an indexed dictionary. Following the match, a

document containing only the match keyword's ids is retrieved. When there are several keywords, only papers containing all of the keywords should be collected for candidate answer passages. This is accomplished by getting the intersections of all document ids containing the keywords. For example, the phrases "duty" and "parent" appear in files 1, 2, 3, 5, and 8, respectively. All files are retrieved as files 1, 2, and 38, and documents matching the keywords in the user query are returned. It can indicate the number of papers that have keyword matches.

**5. Keyword Ranking with Document:**

In the instance of keyword ranking, we first calculate the score between the query terms and all files received through document retrieval. We utilise the Jaccard similarity function to determine the score. In the instance of Jaccard similarity, we first locate the intersection of all files obtained following document retrieval and the extracted keywords of the query.

Score=(A n B)/(A u B)

Were,

**Table1.Keywords** A= set of extracted keywords.

B= set of files keywords

E.g. Input Question- What is the parent's responsibility to ensure their children's education?

Output: Document extracted after jaccard.

**6. Answer Extraction:**

In the case of response extraction, POS tagging is applied to all filter documents obtained by keyword ranking. We examine the sense in between extracted which is obtained by query and filter document after applying POS tagging.

E.G. Assume the extracted query terms are [duty, parent, education, children].

We checked where [Duty, parent] was used as a noun in the document [Education] was used as a

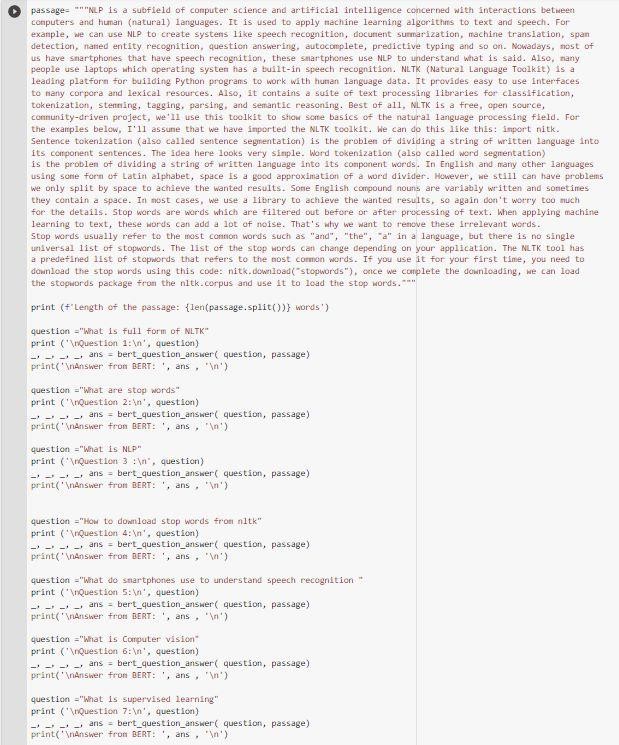
pronoun in the document [Children] was used as a

NNS

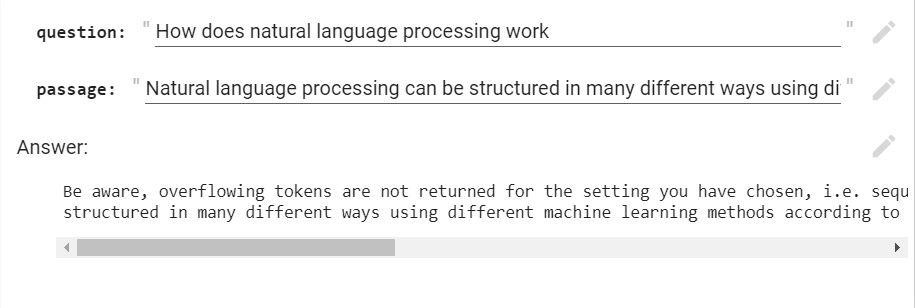
After ensuring that the query and document make sense. We extract the paragraph with the same sense as the query.

**VI. Results and Discussions:**

The use of NLP for question-and-answer applications has yielded encouraging results, with numerous successful applications in diverse disciplines. The IBM Watson system, for example, employs NLP methods to answer natural language inquiries. Watson has been employed in a variety of fields, including healthcare, finance, and education, and has proven to be very accurate and relevant in answering complicated inquiries. Another example is Google Assistant, which use NLP to comprehend user inquiries and offer suitable responses. Users have widely accepted Google Assistant, and it has become an essential part of many people's everyday lives. Despite these achievements, there are still problems that need to be solved in the creation of Q&A apps employing NLP. One significant difficulty is the accuracy and relevancy of the responses produced by these applications. While NLP approaches have advanced greatly in recent years, constructing models that can reliably comprehend and create natural language remains a challenge. Another thing to consider is privacy and security. NLPpowered Q&A apps frequently require access to personal and sensitive data, such as medical records, financial data, and personal preferences. It is critical that these apps include strong security features to preserve user privacy and prevent unauthorised access. Furthermore, while designing Q&A apps with NLP, ethical issues must be taken into account. For example, it is necessary to guarantee that these apps are built and taught in a fair and transparent manner, and that they do not perpetuate prejudices or discrimination against specific groups of individuals.



**Figure4.making of the questions from the passage**



**Figure5.Question and Answer from the passage VII.*Conclusion:***

The purpose of this essay is to look at some of the implementation methodologies and approaches used for Question Answering Systems. We may conclude from a literal survey that an NLP-based question-answering system is more complex than other forms of information retrieval systems. QA systems may be created for resources such as the web, semi-structured knowledge stores, and structured knowledge bases. Despite the fact that this system is limited to a single domain, closed domain QA systems produce more accurate results than open domain QA systems. A closed domain quality assurance system for documents related to education acts is proposed in order to deliver correct and properly more right answers to user concerns.

**VIII.*Future Work:***

There are various potential areas for future research in the field of NLP-based Q&A applications. One alternative option is to incorporate more sophisticated techniques in natural language comprehension, information retrieval, and natural language production to increase the accuracy and relevance of the responses supplied by these applications. Deep learning and reinforcement learning, for example, might be used to increase the performance of NLP models, while more advanced semantic analysis techniques could be employed to better interpret the meaning of user queries. Another potential research path is to improve the user experience of Q&A apps by making them more interactive and personalised. As an example, Chatbots and virtual assistants might be programmed to better grasp the user's preferences, interests, and context, allowing them to give more personalised replies. Furthermore, new modalities such as speech and gesture detection might be included to make Q&A application interaction even more natural and intuitive.

Finally, Q&A solutions that can handle more complicated and nuanced inquiries, such as those requiring reasoning and inference, are needed. Q&A apps, for example, might be developed to answer questions that require many steps or logical deduction, such as "What is the fastest route to get from point A to point B during rush hour?" This would necessitate the creation of more sophisticated natural language comprehension and reasoning models capable of handling more complicated questions.

**References:**

[1]Amit Mishra, Nidhi Mishra, and Anupam Agrawal, "Context-Aware Restricted Geographical Domain Question Answering System", *2010 International Conference on Computational Intelligence and Communication Networks*.

1. Rivindu Perera, "IPedagogy: Question answering system based on web information clustering", *2012 IEEE Fourth International Conference on Technology for Education*.
2. K. Pragisha and P. C. Reghuraj, "A Natural Language Question Answering System in Malayalam Using Domain Dependent Document Collection as Repository", *International Journal of Computational Linguistics and Natural Language Processing*, vol. 3, no. 3, March 2014, ISSN 22790756.
3. Payal Biwas, Aditi Sharam and Nidhi Malik, "A Framework For Restricted Domain Question Answering Sysem", *2014 International conference on issue and challenges in intelligent computing Technique*.
4. Payal Biwas, Aditi Sharam and Nidhi Malik, "A Framework For Restricted Domain Question Answering Sysem", *2014 International conference on issue and challenges in intelligent computing Technique*.
5. B. Katz, "Annotating the World Wide Web using natural language", Proc. of RIAO 97, MCGill University, Canada, 1997.
6. D. Jurafsky and J.H. Martin, Speech and language processing, Prentice Hall, NJ, USA, 2000. [8] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here", Natural Language Engineering, 7 (4), 2001, pp. 275-300.
7. T.R. Gruber, "A translation approach to portable ontology specifications", Knowledge Acquisition, 5 (2), 1993.
8. R. Gaizauskas et al., "The University of Sheffield's TREC 2003 Q&A Experiment", in L.P. Buckland and E. Voorhees (eds): Proc of TREC 2003, NIST, Gaithersburg, USA, 2003.