International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

# SURVEY ON ABSTRACTION BASED RESPONSE SYSTEM USING NATURAL LANGUAGE PROCESSING

**Mrs. M. Therasa [1], A.P. Monisha[2], T. Savitha[3] , S.Sivabharathi[4]**

[1] *Associate Professor, Dept. of Computer Science & Engineering, Panimalar Institute of Technology, Chennai, India*
[2,3,4] *Student , Department of Computer Science and Engineering, Panimalar Institute of Technology, Chennai, India*

---

**Abstract -** *Abstraction based response system is basically summarizing of the given paragraph using natural language processing and machine learning. The amount of text data available has exploded from a number of sources. Hence, this is the procedure for extracting the most crucial information from a source document. This is a common issue in natural language processing and machine learning (NLP). To better help discover relevant information and ingest relevant information faster, an abstraction-based response system is urgently needed to address the ever-growing amount of text data available online. The purpose of the project is to develop a chatbot with abstractive response. A chatbot is a piece of artificial intelligence software that uses messaging apps, blogs, mobile apps, or the phone to simulate natural language communication with a user. Chatbot acts as digital assistant by understanding the human capabilities. Whenever user raises a question the chatbot interprets, process the user request, and gives the relevant answer. In this project the process of abstractive summarization is introduced. It is the process of delivering short and precise answer thereby making the user to understand easily. The generated summaries potentially contain new phrases and sentences that may not appear in the source code by using the concept of abstraction, thereby saves time and effort. Unnecessary data must be reduced to a minimum. Manually summarizing a document is extremely difficult so there is a great need of automatic methods. Approaches proposed is inspired by the application of certain techniques from natural language processing methodology.*

***Key Words*: Abstractive Summarization, Chatbot , Natural Language, Passage Retrieval.**

## 1.Introduction

With the growing amount of data, finding brief information has become difficult. Having a structure that can condense like a person is important in this regard. A synopsis of a given archive is provided by a programmed content rundown with the help of Normal Dialect Handling. There are two types of content outline strategies. i.e. - extractive and abstractive approach.[1]The extractive approach basically chooses the various and unique sentences, sections and so forth make a shorter type of the first report. The sentences are measured and chosen based on the sentences' accurate highlights. In the Extractive technique, we must select a subset from a given expression or set of sentences in a given synopsis frame. The extractive outline frameworks depend on two methods i.e. - extraction and expectation which includes the arrangement of the sentences that are essential in the general comprehension the archive [2]. What's more, the other methodology i.e. producing entirely new articulations to capture the significance of the first record is part of the abstractive content synopsis. This technique is all the more complicated, but it is also the methodology that people use.

The process of creating a description from a given piece of material is a very abstract one in which everybody takes part. Automating such a process will assist in the parsing of large volumes of data and allow humans to devote more time to making important decisions.[3] With so much media

available, it's possible to be very effective by eliminating the fluff from around the most relevant facts. Text summaries created automatically have already begun to appear on the internet. **Abstraction based response system**, thus, is an exciting yet challenging frontier in **Natural Language Processing** (NLP).[4] The latest developments can be traced back to Hans Peter Luhn's paper "The automatic production of literature abstracts," which was published in the 1950s. Our system is implemented with the algorithm natural language processing (Lemmatization, word2vec, N-gram). It is concerned with the interactions between computer and human language. NLP allows computers to read text, listen to speech, interpret it, and decide which parts are essential. A chatbot is a computer program in which conversation occurs in the form of text or audio. The traditional chatbots merely returns extractive summary as the response whereas in this project abstractive responses are implemented.[5] The responses are very clear, easy to understand and thus saves time and effort. This is the key concept of the abstraction-based response system using natural language processing.

### 1.1Proposed System

In our system, the response to the end user is abstractive whereas the bot itself generates the response statement. The question is preprocessed, and keywords are abstracted and then by the keywords the passage retrieval for the topic is done. From the retrieved passage the customized sentence retrieval is made to give a very realistic response to the end user and summarization of the customized sentence is carried forward for the experience of real-world response.

## 2 Methodology

1. Question Processing
2. Passage Retrieval and Sentence Retrieval
3. Answer Processing and Text Summarization

### 2. 1 Question Processing

A Q&A method necessitates a large amount of data and experience, and it is also difficult to incorporate. A chatbot, also known as a Conversational agent, is a service that you communicate with through a chat interface and is controlled by rules or artificial intelligence(little)[6]. Firstly, bot accepts the question from the user and identifies the keywords. bot identifies type of question and searches for the exact match from the Dataset for the data related to the keywords. The Word Embedding technique is used, which is a form of word representation that enables machine learning algorithms to understand words with similar meanings. It entails using a neural network, a probabilistic model, or dimension reduction on the word co-occurrence matrix to transform words into real-number vectors. Word2vec learns words by anticipating their meaning. The syntactic and semantic relationships between words are both encoded. This assists in the discovery of related and equivalent terms. A back propagation method is used to modify or update these features in relation to neighbor or context terms. In NLP, lemmatization is the process of mapping many different forms of the same term to a single form, which we refer to as the root form or base form[7]. The root form is known as a lemma in more technical terms. We reduce our data space and don't have to verify every single form of a word by restricting the number of forms a word can take.

### 2.2 Passage Retrieval and Sentence Retrieval

We achieve passage retrieval by generating question vector and vectors of passage using TF-IDF as feature, which computes cosine similarity between question vector and passage vector returning top 3 closely resembling passage.

The elimination of Stop Words and the use of Porter Stemmer strengthened this step even further. In data recovery, word frequency-inverse document frequency is widely used to handle visit occurring words in a corpus of

related documents.[8] The motivation is to engrave the following question: Are all content terms that appear in documents as often as possible equally important? For example, all records in a collection of news articles investigating the seismic tremor fiasco would clearly contain the word 'quake.' In this way, the tf-idf possibility is to minimize the weight age of visit occurring terms by analyzing their corresponding frequency in the document set.[9] Because of this property, the tf-idf is one of the most widely used terminologies in extractive synopsis. The frequency is defined as:-

$$tf_{j,j} = \frac{n_{i,j}}{\sum n_j}$$

Where:$n_{i,j}$ represents the frequency count of the word i in

document j.

From the total number of words in document j, each word is partitioned and standardized. The term used to weight the calculation is identical to the word likelihood calculation given in Condition 1, and the total number is then divided into various numbers of documents in the corpus that contain different terms.[10] The words I and j are calculated based on Conditions. It tokenizes sentences and computes ngram similarity between the query and the sentence after retrieving the passage. As a result, the most important sentences are identified.

### 2.3Answer Processing and Text Summarization

Module based on the expected answer type, it processes the answer sentence to identify particular entity using name-entity recognition technique and part of speech tagging technique. A natural language processing (NLP) technique for automatically recognizing and categorizing named entities in a text is named entity recognition (NER),[11] also known as entity identification or entity extraction[12]. People, organizations, locations, times, amounts, monetary values, percentages, and other entities are examples of entities. You may use named entity recognition to retrieve key information to find out what a text is about, or simply to gather essential data to store in a database. Machine learning helps machines learn and evolve over time, while NLP explores the structure and rules of language and develops intelligent systems capable of deriving meaning from text and expression. If the question is a description or the bot is unable to recognize the named-entity from the question, it uses the n-gram tilting technique to summarize the document. [13]This module employs the N-GRAM technique. As a result, the terms are modelled such that each n-gram is composed primarily of n -words. An N-gram model, in essence, predicts the occurrence of a word based on the occurrence of its N – 1 preceding word.

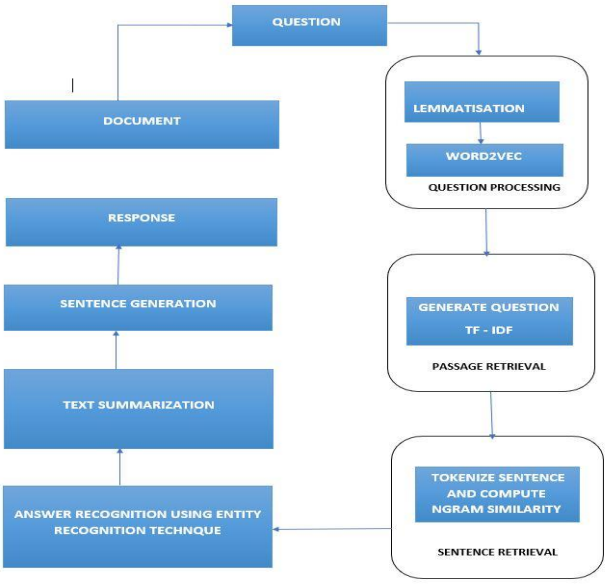


**Fig -1**: Architecture diagram

**Sample Output:**

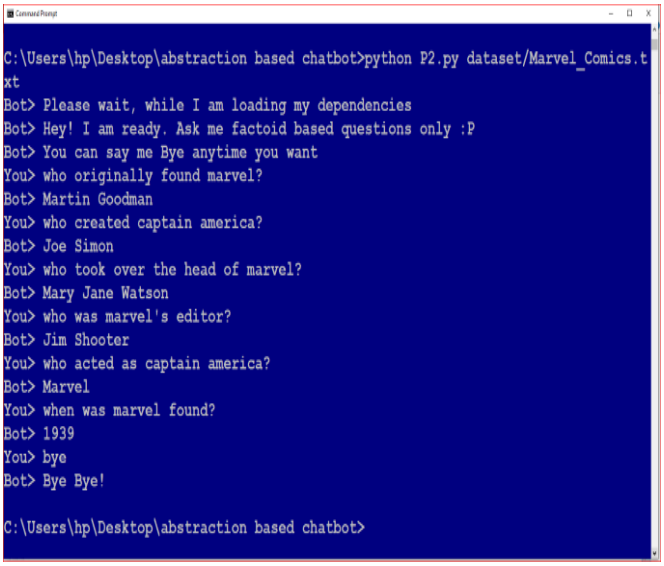


**Fig -2**: sample output answers of a feeded document.

## 3. CONCLUSION

As the internet increases in popularity, data and information is also increasing with-it. Humans would have a difficult time summarizing vast amounts of data. Because of the large amount of data, automated text summarization is needed. We've read a number of articles about text summarization, natural language processing, and lesk algorithms so far. There are a number of automated text summarizers available that have a lot of features and produce good performance. We have learned all the basics of Extractive and Abstractive Method of automatic text summarization and tried to implement extractive one such system which makes use of Natural Language processing technique.[14] The project concludes that this concept of Abstractive Summarization would bring in a revolution in the traditional chatbots and make the people feel a customized response and precise knowledge in the things they are in search off.

## FUTURE SCOPE

We have implemented text summarization using abstractive method. Furthermore, the accuracy of the summarizer after using RNN and LSTM is not satisfactory.[15] Furthermore, we will employ machine learning for semantic text summarization in order to generate more precise summaries, and we will attempt to create a grader that will grade the document using English grammar. There are numerous text summarizers available, but none of them produce appropriate results. As a result, we will use a machine learning algorithm to improve the automated summarizer's effectiveness.

## REFERENCES

[1] Y. Yang W.-T. Yih and C. Meek "WikiQA: A challenge dataset for open-domain question answering" Proc. Conf. Empirical Methods Natural Language Process. pp. 2013-20182015.

[2] S. Balakrishnan A. Y. Halevy B. Harb H. Lee J. Madhavan A. Rostamizadeh et al. "Applying web tables in practice" Proc. Biennial Conf. Innovative Data Syst. Res. pp. 1483-1488 2015.

[3] X. Du J. Shao and C. Cardie "Learning to ask: Neural question generation for reading comprehension" Proc. Annu. Meet. Assoc. Comput. Linguistics pp. 1342-1352 2017.

[4]Q. Zhou N. Yang F. Wei C. Tan H. Bao and M. Zhou "Neural question generation from text: A preliminary study" Proc. Nat. Language Process. Chinese Comput. - 6th CCF Int. Conf. pp.662-671Nov.2017.

[5] P. Domingos "A few useful things to know about machine learning" Commun. ACM vol. 55 no. 10 pp. 78-87 2015.

[6] J. Li W. Monroe A. Ritter M. Galley J. Gao and D. Jurafsky "Deep reinforcement learning for dialogue generation" Proc. 2016 Conf. Empirical Methods Natural Language Process. pp. 1192-1202Nov.2016.

[7]R. J. Williams "Simple statistical gradient-following

algorithms for connectionist reinforcement learning" Mach. Learn. vol. 8 no. 3/4 pp. 229-256 1992.

[8] P. Krishnaveni; S. R. Balasundaram, "Automatic text summarization by local scoring and ranking for improving coherence", International Conference on Computing Methodologies and Communication (ICCMC), 08 February 2018.

[9] C. Lakshmi Devasena; M. Hemalatha, "Automatic Text categorization and summarization using rule reduction", IEEE-International Conference On Advances In Engineering, Science And Management , 2018.

[10] Asha Rani Mishra; V.K Panchal; Pawan Kumar, "Extractive Text Summarization - An effective approach to extract information from Text", International Conference on contemporary Computing and Informatics (IC3I), 06 April 2020.

[11]Shuxia Ren; Kaijie Guo, "Text Summarization Model of Combining Global Gated Unit and Copy Mechanism", IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), 19 March 2020.

[12] Jin-ge Yao, Xiaojun Wan and Jianguo Xiao, "Recent advances in document summarization", *Knowledge and Information Systems*, vol. 53, no. 2, pp. 297-336, 2017.

[13] L. Wang, J. Yao, Y. Tao, L. Zhong, W. Liu and Q. Du, "A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization", *arXiv preprint arXiv:1805.03616*, 2018.

[14] J. Lin, X. Sun, S. Ma and Q. Su, "Global encoding for abstractive summarization", *arXiv preprint arXiv:1805.03989,* 2018.

[15] P. Li, W. Lam, L. Bing and Z. Wang, "Deep recurrent generative decoder for abstractive text summarization", *arXiv preprint arXiv:1708.00625,* 2017.