IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 1, January 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001030 |

Open Domain Question Answering System

Harjeet Kaur Chahal, Dr. Sharvari S. Govilkar

PG Student, Dept. of C.E., Pillai College of Engineering, New Panvel, India

H.O.D., Dept. of C.E., Pillai College of Engineering, New Panvel, India

ABSTRACT: Technology has advanced and hence people in recent time's uses internet search engines as basic medium for understanding of their concept and knowledge level doubts. The whole idea is to provide an architecture which is the healthy way of searching for the information available on the internet and it is based on the cognitive computing. Our system is intended to manage and mining suitable knowledge from large amount of textual data. Thus resulting in providing exact answer to the natural language query asked by the user with a new level of refinement to search results for the users. The idea is to propose a system to tackle open domain question answering system using the unique knowledge base of Wikipedia. The text span in a Wikipedia article is the answer to any factoid question. The idea is to consider the problem of answering factoid questions using Wikipedia as the unique knowledge base in an open-domain setting, such as anyone does when searching for answers in any of the encyclopedia. If they are able to leverage its power, that could facilitate intelligent machines as Wikipedia is an evolving as a source of huge information. Unlike other knowledge bases such as Freebase or DB Pedia, which are compared to Wikipedia are easy to process but too sparsely populated for any open-domain question answering. Wikipedia contains Wikipedia contains latest valuable information that humans are interested in.

KEYWORDS: Question Answering system; Neural Network; RNN; Ranking Algorithm; Prediction.

I. INTRODUCTION

Question Answering denoted as (QA) is a fast growing research area and commercial interest. The problem of QA is to find answers to open-domain questions by searching a large collection of documents. QA systems basically provide short and relevant answers to questions. Open domain question answering is an old unsolved problem and rarely explored. Such system requires broad knowledge to ensure time complexity and high coverage at which the data can be analyse and understood by systems. The required information is present mostly in the emails, writer's blogs, articles, forums and research papers. But this data is least useful until it is turned into meaningful information or knowledge [1]. So the main challenge is the increasing complexity of such content provider systems with the actual data volume explosion in real world, putting these knowledge into solving various important issues.

To address above scenario, our open domain QA system is a solution towards automating question answering system which can answer natural language questions. The whole idea is to provide a blueprint which is the new and unique way of searching for the important information available on the internet and it is based on the cognitive computing [2]. Cognitive computing mainly resolves the situation where uncertainty and ambiguity is there and attempts to copy the mechanism of human brain. The problem of answering factoid questions in an open-domain setting using Wikipedia as the unique knowledge base in an open-domain setting, such as anyone does when searching for answers in any of the encyclopedia [3]. If they are able to leverage its power, that could facilitate intelligent machines as Wikipedia is an evolving as a source of huge information. Unlike other knowledge bases such as Freebase or DB Pedia, which are compared to Wikipedia are easy to process but too sparsely populated for any open-domain question answering.

Wikipedia contains latest and updated valuable information that humans are interested in. Using available Wikipedia articles as our knowledge base causes the work of QA to combine the challenges of both machine comprehension of text and large scale open domain QA. In order to provide an answer to any specific question, one must first fetch the some relevant articles from more than 10 million items to identify the answer by scanning them carefully. We can say this as machine reading at scale i.e. MRS. Our task treats Wikipedia as only collection of different articles and also not bother about its internal graph structure. As a result, it makes our approach generic and made us easier to switch to another available collections of documents or daily updated newspapers.

II. RELATED WORK

Ruby Bhati and Prof. S.S Prasad [1] have focuses on development of QA engine using available open source software. It mainly uses cognitive computing as this is one of the emerging paradigm which can mimic the mechanism of the brain by using intelligent computing system. A question answering system can be considered in detail as a prototype and its features can be compared with the features provided by a cognitive systems.Prakash Ranjan and

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001030 |

Rakesh Chandra Balabantaray [2] explains mainly consist of three steps. Question classification which is done to find the type of answer required by a given question. After that data collection in which they fire whole question as a query to the search and extract top 8 retrieved page. For this they have taken the help of Goggle search engine. And finally answer candidate extraction will be done.Zhonglin Ye, Zheng Jia,Yan Yang and Hongfeng Yin [3] organized and summarized the recent research evidences altogether in a new way that integrated understanding of working in the question answering domain. It focuses on the classification of the available literature, developing its own perspective on the area, and evaluating trends out of it on the research area. However, as it is impossible for any survey to conclude based on limited time to include most of previous research. In this survey, they have included only the work of the top-cited/ top-publishing authors in the QA area. As scientific research is nonstop process, this survey also included accumulative activity research containing their limitations to present how these limitations were discovered, faced and treated by other researchers.

Darshana and Nivid [4] mainly contributed in proposing a simple yet effective deep learning methods like transformed phrase into word vector the usage of one-of-a-kind word embedding algorithms. Five phrase embedding fashions are used like Word2vec, Fasttext, Glove, SL999, Baroni. They adopted a combined hybrid approach that involved different complementary components which includes information retrieval, extraction and linguistic tools such as name finding. Also includes parsing, reference resolution. It also includes proposition, relation and extraction of structured patterns. Versatile global T-max pooling is used to extract function based totally on the maximum cost and also used for prediction purpose to predict the subsequent word in the collection. This paper focusses on predicting high-quality answer among all solutions and fed it for ranking purpose.Adam, Jason, Antoine and Danqi Chen [5] explains datasets which have been used like Wikipedia that they have used as knowledge base for finding answers to the queries, the SQuAD dataset act as a main resource to train the Document Reader and three more datasets that in addition to SQuAD they have used to test the open domain QA performance of their full system. There experiments show that the Document Retriever they have used to retrieve the documents has outperforms the Wikipedia search engine. Document Reader reaches the results on the very competitive benchmark of testing results in improved amounts.

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- In this approach system will retrieve the correct answers to the queries posed by the user in our search engines. Our proposed system will provide the require answers from Wikipedia source. The system architecture is shown in Fig 1.
- Input Open Domain Questions should consider the Un-normalized input.
- Pre-processing needs to be done.
- Keyword Search module to identify the keywords.
- Algorithms needs to be applied.

B. *Description of the Proposed Algorithm:*

Step 1: Input Open Domain Questions:

This module will take the input as single word, phrase or sentence. We will validate the same using question type system and ensure the length of the question should reside in mentioned options only.

Step 2: Pre-processing:

Pre-processing involves the below steps which will ensure some important points like To extract useful information from the unstructured text, several pre-processing steps are applied to remove spelling errors, grammatical mistakes, etc.

Student feedback data represents an unstructured text. Sentence is segmented by identifying the boundary of sentence which ends with full stop. It will take the first sentence only from the input if we have more than one sentence found in the input.

Tokenization is the process of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements words by identifying the comma, spaces and special symbols between the words called tokens. Tokens can be individual words, phrases but not a whole sentences. In the process of tokenization, some characters like punctuation marks are discarded.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001030 |

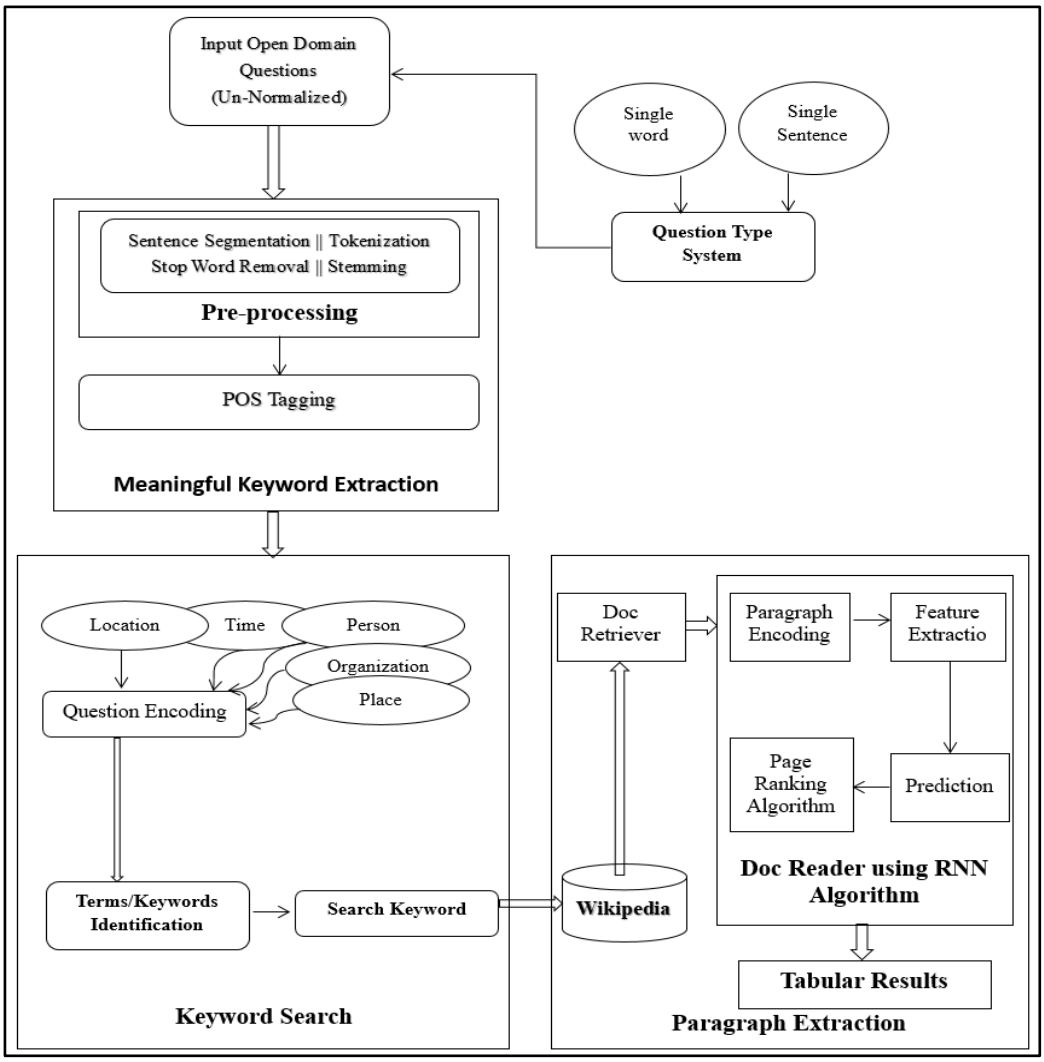

Fig.1. Proposed System Architecture

Stop words are used in every language. We need to only consider the words in the document which have importance to effectively use word feature score. Stop words are eliminated to focus on important words. For example, search engine query is “Who is the Prime Minister of India.”, in this type of query search engine searches for the words like “who”, “is”, “the”, ” Prime”, ” Minister”, ”of”, ”India”. Then it retrieves more pages containing the words “India”, “Prime”,” Minister”. So, by deleting or removing these stop word we can get proper data for analysis. The words are reduced to its base form or root word which is known as stem. Stemming is a mandatory pre-processing step in number of natural language processing applications like word sense disambiguation. One of the widely used tools for this processing is stemmer which keep a suffix list to remove suffix from its word. Porter stemmer which is one of the popular stemming algorithm is a process for removing the commoner morphological and in flexional endings from words in English. It removes prefixes and suffixes to get the affix free word by porter rules.

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that assigns parts of speech to each word by reads text in some language, such as noun, verb, adjective, etc. It is a very strong and useful tool. It is used in such an application that deals with text to analyse words or tokens and classify them into different categories. Our POS tagging tool is uses a probability model. It predict out the correct POS tag from the give tag set. We used WordNet parser to parse the sentences and WordNet tagger for the part of speech tagging. WordNet dependencies were utilized to extract the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001030 |

sentence syntactic elements (subject, verb, object) also. We uses Part-of-speech tags as it is used in Penn Treebank Project

C. *Keyword Search:*

First step is question encoding where we have to check if input contains any location, time, organization, person or place by giving the tags based file which contains the location, time, organization, person or place as input to the module.

Terms/Keywords Identification is a next step where we check the meaningful keyword identified or not and check for those keywords on Wikipedia.

Doc Retriever is used as we are deciding the number of documents to show to the users in the end, Default value will be 5 docs.

D. *Algorithms:*

First step is Paragraph Encoding as we first represent the all available tokens pi in a available paragraph p as a of feature vectors of sequence P1 ϵ Rd. Then pass them as the input to RNN i.e. recurrent neural network and thus obtain.

$$\{p1\ldots pm\} = RNN\ (\{P1\ldots PM\}) \qquad \text{eq. (1)}$$

Where pi is expected to encode useful context information around token p. Here, we will choose to have a long short-term memory network which will be multi-layer bidirectional and take pi as the concatenation of each layer's hidden units in the end. The feature vector P1 is comprised of the following parts, we suppose to use three binary features, indicating whether pi can be exactly matched to one question word in q, ignoring the case sensitiveness of the word.We also need to add a few extra features which shows some properties of token pi, which include the part-of-speech (POS) of the word, named entity recognition (NER) tags and its term frequency (TF) to be precised.

Feature extraction phase will use RNN using LSTM, the tokens which was identified will be in memory to have the prediction step more accurate and page rank will provide the order as per the weighted as per Token found in paragraph.

At the paragraph level, the goal is to predict the span of tokens that is most likely the correct answer. We take the paragraph vectors {p1…….pm} and the question vector q as input, and simply train two classifiers independently for predicting the two ends of the span. Concretely, we use a bilinear term to capture the similarity between pi and q and compute the probabilities of each token being start and end as

$$Pstart\ (i)\ \alpha\ exp\ (pi\ Ws\ q)$$
$$Pend\ (i)\ \alpha\ exp\ (pi\ We\ q) \qquad \text{eq. (2)}$$

During prediction, we choose the best span from token i to token i' such that i <= i' <= i + 15 and Pstart (i) * Pend (i) is maximized. To make scores compatible across paragraphs in one or several retrieved documents, we use the normalized exponential and take argmax over all considered paragraph spans for our final prediction

Page Ranking Algorithm, here we will rank the paragraph output form prediction step using term frequency using below formula,

Term Frequency = (Number of times term t appears in a document) / (Total number of terms in the document). eq. (3)

IV. CONCLUSION AND FUTURE WORK

From the research it can be concluded that automatic QA system based on domain knowledge base can directly get an intuitive and accurate answer, so it becomes an important focus of research. We are providing an enhanced solution for open domain QA system where we are focusing on Document summary and using neural network to provide the solution to any domain question. We will be using Wikipedia as the source of information for providing the answer to the user queries.

REFERENCES

1. Ruby Bhati, Prof. S.S Prasad , "Open Domain Question Answering System using Cognitive Computing", 6th International Conference - Cloud System and Big Data Engineering (Confluence), 2016.
2. Prakash Ranjan, Rakesh Chandra Balabantaray, "Question Answering System for Factoid Based Questions", 2nd International Conference on Contemporary Computing and Informatics (ic3i), 2016.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001030 |

3. Zhonglin Ye, Zheng Jia,Yan Yang and Hongfeng Yin , "Research on Open Domain Question Answering System ", Springer International Publishing Switzerland, 2015.
4. Darshana V. Vekariya, Nivid R. Limbasiya,"A Novel Approach for Semantic Similarity Measurement for High Quality Answer Selection in Question Answering using Deep Learning Methods", 6th International Conference on Advanced Computing & Communication Systems (ICACCS), 2020.
5. Adam Fisch, Jason Weston & Antoine Bordes, Danqi Chen, "Reading Wikipedia to Answer Open-Domain Question", Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017.

BIOGRAPHY

Harjeet Kaur Chahalis a Research Assistant in the Computer Engineering Department, Pillai College of engineering, Mumbai University. She received Bachelor of engineering (B.E.) degree in 2015 from Mumbai University, India. Her research interests are Computer Networks (wireless Networks), Machine Learning and Algorithms, etc.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details