International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1471-1473 www.ijaem.net ISSN: 2395-5252

Searching for Relevant Text

Aman Jaglan*1, Dr.Sunil Maggu*2
Student, Information Technology, Maharaja Agrasen Institute Of Technology, Rohini – Delhi, India.
AssistantProfessor,InformationTechnology,MaharajaAgrasenInstituteOfTechnology, Rohini, Delhi, India.

Submitted: 15-05-2022 Revised: 20-05-2022 Accepted: 25-05-2022

ABSTRACT
The area of material and knowledge on any given topic of interest is always expanding. This widening of a subject's knowledge base, along with simple access to the Internet via various devices, has resulted in quick access to a wealth of material when conducting a search relevant to the topic. However, this massive data set also poses a hurdle. To find the needed information, one must put out some effort. An individual may be forced to go through a number of documents in order to locate the answer that they are looking for.
This work focuses on a model for dealing with this problem that aids in the discovery of a relevant and exact solution to a question. The model combines Natural Language Processing algorithms with knowledge graphs to achieve this goal. The study comes to a conclusion by detailing the model's performance.
Keywords: Natural Language Processing, Entity Extraction, Knowledge Graph, Parsing, and Question-Answering.

I. INTRODUCTION

Students today are surrounded by vast amounts of knowledge available on the Internet in the form of full text papers. However, in most cases, a student would like to have a straight and specific solution to his or her inquiry rather than sifting through a lengthy text to discover the proper answer. Modern search engines such as Google, Microsoft Bing, and others offer impressive capabilities, however they generally present a list of documents or web pages linked to a user's query. So, in order to obtain the solution, one must read the entire document. It's also possible that a paper provided has a lot of information concerning the subject sought but excludes the precise information that the user is looking for

For instance, consider the question "Which vitamins are found in carrots?"

Users would be more interested in direct responses such as Vitamin A, C, and K rather than a big page that they must read to locate the required answer.

Using Natural Language Processing (NLP) techniques and knowledge graphs (a term popularised by Google in 2012), this article makes it easier to get direct answers to such inquiries if the required language or information is available. Since then, Google's knowledge graph (KG) has grown at an exponential rate. As of May 2020, Google's database comprises roughly 500 billion information relating to 5 billion entities..

II. RELATED WORK

This field has seen a significant quantity of research in the past. The following sections cover some of the prior work in this topic.

In [1,] a community QA system is used to score several assertions (comments) in order of excellent, bad, and possible replies when given multiple statements (comments) connected to a question. Convolution neural networks (CNN) with convolution filters 3,4,5 are used in the model to achieve this. Furthermore, 100 smart neurons are employed as long term short memory (LSTM). Finally, the most appropriate reply is given as an answer to the query. With only CNN, this model has an accuracy of 68.8%, 66.7 percent with only LSTM, and 72.5 percent with both CNN and LSTM.

In article [2,] the author proposes a closed domain question-answering system based on educational laws. Data about educational acts is gathered and pre-processed using stemming and stop word removal. The extracted keywords are saved in an index term dictionary with the keyword as the key and a list of document numbers that include the associated keyword. The intersection of document numbers of all the matching keywords is extracted and scored using the jaccard similarity function for a query.The papers are then tagged with part of speech (POS) tags, and the document with keywords that have the same grammatical sense as the query's keywords is returned as the final result. As previously stated, this approach is limited in that it only delivers the entire page as an answer rather than a straight response to the question.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1471-1473 www.ijaem.net ISSN: 2395-5252

III. METHODOLOGY

We will offer an overview of data collection and transformation in this part, which will serve as the foundation for our structured dataset for the question answering system and knowledge graph. We'll also show you how to use a high-precision query retrieval technique.

A. Data acquisition

Despite the fact that our suggested method works with a wide range of texts, we picked textbooks from courses such as geography and history to demonstrate how it may be used in practise. Our primary data source is kaggle, from which we will obtain various educational texts on the topics of our choice (mainly theoretical subjects). Each subject was examined for about 500 pages. The data is originally in unstructured text format, which is not ideal for our immediate usage. As a result, the next step is to preprocess the data and make it usable. We need to store the quintets for each phrase now that we have them in the form of a dataframe.

B. Using JSON to store entities and relationships

The resulting quintets must be saved in an organised style. We choose json as the storage format out of numerous possibilities such as csv, xml, and others. These quintets will be utilised for constructing the knowledge graph and searching the dataset.

C. Creation of knowledge graph

A knowledge graph is a way of displaying data by integrating it into an ontology, and it aids the system in deriving new knowledge from the graph. It is employed because it is a very flexible data structure that allows us to dynamically map our json file to the graph. The graph is shown using the Python NetworkX module. Each quintet's subject, object, and predicate are utilised to build a graph edge**.**

IV. RESULT AND DISCUSSION

This section contains the acquired results as well as the different information linked to the experiment. The results are satisfactory, and the strategies employed here were effective for our system.

1. The Implementation Hardware

The algorithms are implemented using the setup shown below. All of this hardware is utilised in the implementation. Other hardware can be utilised for implementation as well. Python programming was carried out using a Windows 10 64-bit operating system. NumPy, Pandas, Matplotlib, SciPy, Scikit-Learn, PyTorch, Seaborn, Plotly, TensorFlow, Keras, and Seaborn were used in the implementation**.**

2. The Dataset

The experiment makes use of the SQUAD dataset. This dataset is used in the factoid QA experiment. The dataset contains a range of articles on various issues, as well as a variety of topics, allowing for proper analysis and testing on a number of problems.

3. Passage Retrieval Results

When evaluated over 422 articles from the SQUAD dataset, the accuracy of passage retrieval was 69.69 percent. This accuracy measure increased to 77.49 percent after deleting stop words and applying Porter Stemmer. According to further investigation, the corresponding paragraph appears in the top three returned paragraphs in 94.23 percent of passage retrieval.

1. For accurate responses, user input may be used to increase the accuracy of our suggested method.

2. More complicated queries can be answered by extracting as much data from the context as possible in the form of entities.

3. Improved search algorithms can be employed to provide correct responses in a timely manner.

4. The gathered data may be utilised to train and create a neural network to improve the system's accuracy.

5. After achieving specified degrees of development, the system may be utilised in the classroom to answer students' questions and retrieve specific answers.

V. CONCLUSION

One of the most significant systems for responding consumer inquiries is quality assurance. The factoid QA system handles a wide range of articles and performs a wide range of QA. The provided approach also works for a variety of QA, including lexical chain, NLP, AI, and keyword analysis as methodologies. The SQUAD dataset, which comprises a variety of publications, is used here. The system checks a big number of queries and our experiment is tested on a range of articles. Indore, Bhopal, Durg, and Jabalpur are among the items examined. The system's accuracy is on average 69.93 percent. The factoid QA system may be implemented using the presented system.

REFRENCE

[1]. M. Wakchaure and P. Kulkarni, "A Scheme of Answer Selection In Community Question Answering Using Machine Learning Techniques," 2019 International

Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 879-883, doi: 10.1109/ICCS45141.2019.9065834.

[2]. S. P. Lende and M. M. Raghuwanshi, "Question answering system on education acts using NLP techniques," 2016 World Conference on Futuristic Trends in Research & Innovation for Social Welfare (Startup Conclave), Coimbatore, 2016, pp. 1-6, doi: 10.1109/STARTUP.2016.7583963.

[3]. T. Dodiya and S. Jain, "Question classification for medical domain Question Answering system," 2016 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Pune, 2016, pp. 204-207, doi: 10.1109/WIECON- ECE.2016.8009118.

[4]. Y. Li, J. Cao and Y. Wang, "Implementation of Intelligent Question Answering System Based on Basketball Knowledge Graph," 2019 IEEE 4th Advanced Info Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019, pp. 2601- 2604, doi: 10.1109/IAEAC47372.2019.8997747.

[5]. https://youtu.be/zeYfT1cNKQg

[6]. https://ieeexplore.ieee.org/document/7947211

[7]. A Specialized Question Answering System Using NaturalLanguage: Developed Using the Quepy Framework

[8]. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/253938

[9]. https://rdflib.readthedocs.io/en/stable/