Journal of Advanced Zoology
ISSN: 0253-7214
Volume 44 Issue 5 Year 2023 Page 605-611

Interactive Malayalam Question Answering System: A Neural Word Embedding And Similarity Measure Based Approach.

Liji S K[1*], Muhamed Ilyas P[2]

[1*,2]*Sullamussalam Science College, Malappuaram, Kerala.* [1]*E-mail: liji.s.k@gmail.com*

***Corresponding Author:** Liji S K*
**Sullamussalam Science College, Malappuaram, Kerala. E-mail: liji.s.k@gmail.com*

Article History

Received: 08/09/2023
Revised: 12/10/2023
Accepted: 16/11/2023

Abstract

This innovative system operates as an automated, domain-specific knowledge repository designed specifically to furnish reliable Malayalam responses to inquiries pertaining to COVID-19. Leveraging advanced Natural Language Processing (NLP) algorithms, both Malayalam documents and questions undergo meticulous processing. The semantic modelling and document conversion stages employ the Word Embedding approach, specifically Continuous Bag of Words (CBOW), to enhance the system's understanding of the language nuances. Subsequently, the retrieved results for a given query are meticulously ranked using the cosine similarity measure, ensuring that the most relevant and accurate information is presented to the user. Integral to the system's efficacy is our proprietary Malayalam question-answering dataset. This dataset has been meticulously curated, drawing from reliable and publicly accessible sources related to COVID-19. It serves as the foundation for experimentation, reflecting the system's ability to provide accurate responses. The system's performance is quantified using the F1 score, a metric that combines precision and recall, yielding a comprehensive evaluation. In our experimentation, the F1 score of the Semantic Malayalam Question-Answering System is found to be 76%, attesting to its robustness and effectiveness in delivering trustworthy information in the Malayalam language within the context of COVID-19.

Key words: Question Answering; Natural Language Processing; Continuous Bag of Words; Information Retrieval; Cosine Similarity.

1. Introduction

The severe acute respiratory syndrome (SARS) that is the cause of the global coronavirus pandemic COVID-19 originated in China. A sudden COVID-19 outbreak forced millions of individuals into quarantine and social distance. As a result, reports from all across the world indicated a sharp increase in mental health issues among people, including depression, stress disorders, and suicidal thoughts. As a result, individuals are always searching the internet for COVID-19 updates and solutions. A plethora of web sites are at one's disposal for obtaining knowledge.We suggest a Malayalam question-answering system that can automatically respond to COVID-19-related queries in order to address this problem. Contextual and semantic information are required for the semantic information retrieval from vast amounts of unstructured data. We suggested a word embedding-based question answering system to bridge the semantic gap.

This is an automatic domain-specific knowledge system designed to provide pertinent Malayalam responses to queries pertaining to COVID-19. The Malayalam language is used by users to query the system, which will then process their questions, compare them with the documents that are available, and rank and get the

responses that are most pertinent to their queries. While there are numerous intelligent systems available in other languages, the Malayalam language does not currently have an efficient system for answering semantic questions. Developing a semantic Q&A system is a difficult endeavor because of the agglutinative character of the Malayalam language. Even so, the Q&A feature in our own tongue will benefit native speakers, particularly the illiterate ones. This work's primary motivation is this.

2. Question Answering Systems

Question-Answering and Information Retrieval are utilized to automatically obtain pertinent responses in natural languages for consumers' inquiries. The technology allows users to query it in their own language. After processing the queries and matching them with the documents, the system will return the pertinent results. Native languages like Malayalam, Kannada, and others can be processed and analyzed using natural language processing techniques.
Users submit natural language questions to the question-answering system, which then interprets and transforms them into more structured or meaningful forms. After that, the algorithm classifies them by analyzing the organized shape. Next, compare the queries to documents that are already in the database, assign a score to the documents based on any similarity metric, and obtain the relevant result. Figure 1 depicts the general architecture of the question-answering system.

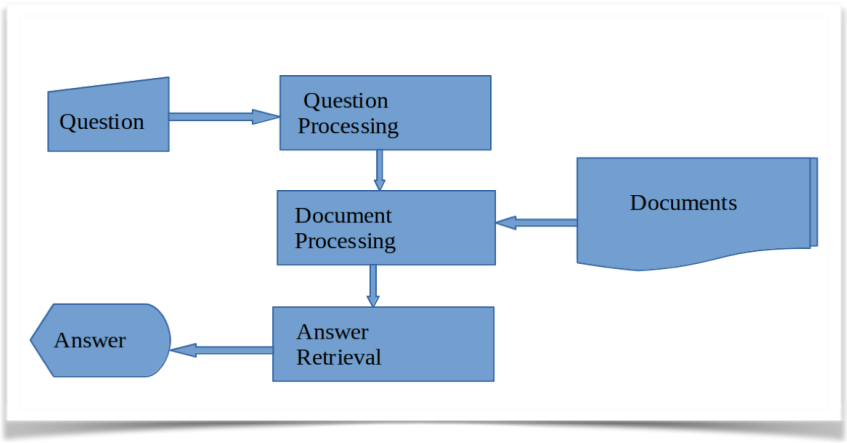

Figure 1: Architecture of Question Answering System

3. Review of Related Work

In a work, conducted by Fan Fang, Bo-wen Zhang, and Xu-cheng Yin as documented in their work [5], a novel approach to information retrieval was introduced. Their methodology involved a synergistic integration of semantic information with the traditional Sequential Dependence Model (SDM). The aim was to enhance the precision and effectiveness of information retrieval processes. A noteworthy aspect of their approach was the incorporation of word embedding-based synonym mapping. This technique is rooted in the utilisation of word embeddings, specifically employing methods like Continuous Bag of Words (CBOW) or Skip-Gram, to map synonymous terms or phrases within the document corpus. By establishing semantic connections through word embeddings, the researchers sought to refine the intricacies of the Sequential Dependence Model.
The Sequential Dependence Model, a conventional approach in information retrieval, relies on the inherent sequential dependencies within a document. It considers the order and arrangement of words to infer context and relevance. By complementing this model with semantic information derived from word embeddings, the researchers aimed to enrich the system's understanding of the relationships between words and phrases, thereby improving the accuracy of the information retrieval process. The use of word embedding-based synonym mapping, in particular, adds a layer of nuance to the retrieval process. It enables the system to recognise not only direct lexical matches but also semantically related terms, thereby broadening the scope of relevant information retrieved. This innovative integration of semantic information with the Sequential Dependence Model reflects a forward-thinking approach to information retrieval, embracing both traditional models and contemporary techniques. The work by Fang, Zhang, and Yin contributes to the ongoing exploration of advanced methods to optimise the extraction of pertinent information from vast document corpora.

Bo Xu, Hongfei Lin, and Yuan Lin [6][7] presented a query expansion on learning-to-rank techniques biomedical information retrieval system. In addition, they employed query expansion, pseudo relevance feedback, and co-occurring term selection, among other techniques.
Researchers Dwaipayan Roy, Debasis Ganguly, Sumit Bhatia, Srikanta Bedathur, and Mandar Mitra [8] all looked into how word normalization and collection decisions affected the performance of ad hoc retrieval when learning word embedding. Two criteria were put forth by them to gauge how similar word vector embedded spaces are to one another. Because of the intrinsic properties of the embedding algorithms, word2vec (1) typically performs well on a stemmed collection and (2) quickly produces text on an unprocessed collection.
Md Mahadi Hasan Nahid, Shomi Khan, and Khadiza Tul Kubra [9] suggested a Bangla question-answering system. They experimented with the technique using both English and Bangla. They matched semantics using anaphora-cataphora resolution.
A technique for answering Malayalam questions was proposed by Archana S.M., Naima Vahab, et al. [10]. They determined the terms in question and the vibhakthi connected to the response. The system is governed by rules. Based on the vibhakthi of the question and response modules, the answers are obtained.
Using query expansion techniques, Arjun Babu and Sindhu L 11] have created an information retrieval system for Malayalam. For answer retrieval, they employed vector space modeling and cosine similarity. Additionally, they employed synonym mapping to raise the system's accuracy.
Vaishali Singh et al. [12] presented a customized method of question answering through end-user modeling in their work. based on the user's profile and region of interest. Several similarity metrics, including entity value similarity and attribute value similarity, can be used to personalize the retrieved data.
A unique technique for utilizing a graph model to determine document similarity was provided by Sheetal.S et al. [13]. They used Wikipedia and WordNet with a modified approach. The weighted conceptual graph of the coexistence term in this method is utilized to display textual documents. To determine the relationship between feature terms and weighted terms for graph construction, they employed the co-reference resolution method.
An ontology-based information extraction and summary system for Tamil news material retrieval for user queries was proposed by Swathilakshmi Venkatachalam et al. [14]. Certain information can be extracted from natural language and submitted to ontology by employing information extraction. There are two distinct domains within the ontology cluster. A multi-document text summarizer then compiles the most significant events into a summary. In the end, the query extractor pulls information from the database and presents it to users based on their search terms.
In the field of music, Navjot Kaur et al. [15] created a semantic information retrieval system. They developed query reformulation techniques in order to perform multilingual information retrieval using string ontology for semantic information retrieval.
According to the review, the majority of semantic information retrieval work is done in the English language domain, while a small number of works are done in native languages like Tamil and Malayalam. Various techniques such as Natural Language Processing (NLP), Machine Learning, Deep Learning, etc. are employed for retrieving information at the semantic level. Word embedding at the context level and deep learning are the foundations of most current efforts.

4. Architecture of Proposed Question Answering System.

Let us explore the details of the suggested semantic Malayalam question answering system, looking at its architecture and the approaches used at different phases. The architecture of the system is made up of a number of essential parts, each of which is vital to maintaining the effectiveness and precision of the information retrieval process. Document ranking, feature vector development and semantic modelling, query and document preprocessing, and response retrieval are some of the crucial phases.

a) Preprocessing of Query and Documents.

Initially, the query is transformed into a standard format by comparing its structure with the queries that have been saved. Next, a variety of pre-processing methods are applied to the query and documents, including lemmatization, stop-word removal, and tokenization.The user requests are divided into discrete tokens throughout the tokenization process. We're using word tokenization here. Next, in order to decrease the dimension of the term space, stop words were removed. In this effort, we create our own list of Malayalam stop words that are not as crucial to the process of answering questions.Lastly, based on the word's intended meaning, lemmatization was employed to determine the word's "lemma." Lemmatization, as opposed to

stemming, aids in determining a word's Part Of Speech, meaning, and context within a document. Here, we carried out all of the preprocessing using iNLTK.

b) Feature Vector Creation and Modelling.

Model the un labeled word from the corpora into a dense vector space after pre-processing. In this case, we suggested modelling the words to an m-dimensional vector space using word embedding. Semantics and word context will be taken into account while modelling. Model a word in word embedding by mapping it from a multidimensional space to a continuous vector space with a significantly less dimension per word. Word2Vec is the algorithm suggested for this investigation [4][5]. Word2vec uses a three-layered neural network to train itself to construct vocabulary from the corpus and learn word representations. Word2vec offers two models: 1. Skip-gram and 2. Continuous Bag of Words (CBOW). Unlike skip-gram, which learns representations by guessing each target word based on its context words, CBOW learns representations by predicting the target word. This study conducted with CBOW model.

c) Document Modelling and Answer Retrieval

The similarity measure, which might indicate how similar the searches and documents are, is used to rank the documents. For answer retrieval, we employ cosine similarity in this case. Equation 1 illustrates the calculation of the Cosine similarity between the questions and the document.

Cosine-sim (Q,D) = [|Q∩D| / |Q|1/2 * | D| 1/2] - (1)

A higher Cosine value suggests that the query and the documents are highly comparable. In response to the user's query, the system will return the document with the highest cosine similarity value.

The block diagram of proposed question answering system is shown in figure 2.

The different steps of the proposed algorithm as shown below.

step1:Start

step 2: Pre-processing of Query and documents.

Step 3: Convert and Model the documents in to a M -dimensional vector space using CBOW Word2Vec algorithm.

step 4: Find the similarity between query and documents by using cosine similarity measure.

Step 5: Rank the documents based on their cosine similarity value. step6: Return the document with the highest cosine similarity value. step7: Stop.

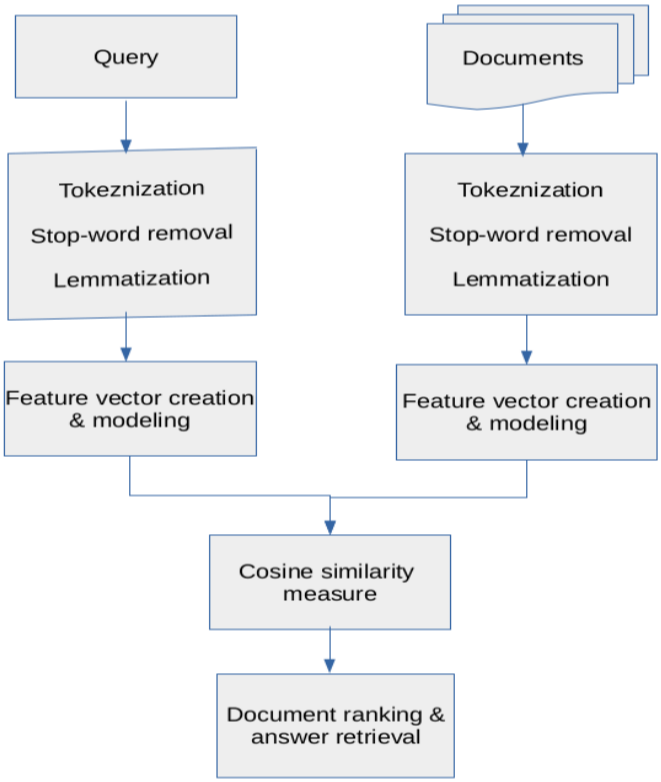

Figure 2 : Block Diagram of Question Answering System.

5. Experimental Results and Analysis

The implementation of the proposed semantic Malayalam question-answering system necessitates the creation of a robust and domain-specific dataset. To achieve this, we have meticulously curated our own Malayalam question-answering dataset by extracting information from publicly available websites such as TDIL (Technology Development for Indian Languages), Kerala.gov.in, health.gov.ac, and others. This dataset forms the foundation for training and evaluating the system's performance.

The process of dataset creation involves systematically collecting relevant textual data, including questions and corresponding answers, from authoritative sources related to COVID-19. The chosen websites are reputable platforms that disseminate accurate and up-to-date information in Malayalam, ensuring the reliability of the dataset. The questions are formulated to cover a diverse range of topics within the domain of COVID-19, reflecting the varied information needs of users.
Here is an excerpt of sample queries and their corresponding retrieved results from the created dataset, as illustrated in Table 1:

NO	Question	Answer
1	കൊറോണ വൈറസ് രോഗലക്ഷണങ്ങൾ എന്തെല്ലാം ? [What are the symptoms of coronavirus?]	പെട്ടെന്നുണ്ടാകുന്ന പനിയും ചുമയും അല്ലെങ്കിൽ പനി, ചുമ, പൊതുക്ഷീണം, തലവേദന, പേശീവേദന, തൊണ്ടവേദന, നാസാഗഹ്വരത്തിലെ വീക്കം, ശ്വാസംമുട്ടൽ, ഛർദ്ദി, വയറിളക്കം, മാനസികാവസ്ഥയ്ക്കുള്ള മാറ്റങ്ങൾ , **എന്നിവയിൽ ഏതെങ്കിലും മൂന്നുലക്ഷണങ്ങൾ ഉൾപ്പെട്ടിരിക്കുന്നു.**[Sudden onset of fever and cough or fever, cough, general fatigue, headache, muscle aches, sore throat, nasal congestion, shortness of breath, vomiting, diarrhoea, and mood swings, includes any of these three symptoms]
2	കൊറോണ വൈറസ് രോഗം പകരുന്നത് തടയാനുള്ള മാർഗ്ഗങ്ങൾ എന്തെല്ലാം ? [What are the ways to prevent the spread of coronavirus disease?]	വ്യക്തിശുചിത്വം പാലിക്കുക, രോഗബാധിതരിൽ നിന്ന് അകലം പാലിക്കുക, ഹസ്തദാനം ഒഴിവാക്കുക, കൈകൾ ഇടയ്ക്കിടെ സോപ്പ് ഉപയോഗിച്ച് 20 സെക്കന്റോളം നന്നായി കഴുകുക, ആൾക്കൂട്ടം ഒഴിവാക്കുക എന്നിവ രോഗപ്പകർച്ച തടയാൻ ശുപാർശ ചെയ്യുന്നു. ചുമയ്ക്കുമ്പോൾ മൂക്കും വായയും മൂടുന്നതിലൂടെ രോഗാണുവ്യാപനം കുറെയേറെ തടയാം. **[It is recommended to maintain personal hygiene, avoid contact with infected people, avoid shaking hands, wash hands thoroughly with soap for 20 seconds frequently, and avoid crowding. Covering the nose and mouth when coughing can help prevent the spread of the disease.]**
3	പ്രായമായവർക്ക് കൊറോണ വൈറസ് രോഗം വരാൻ സാദ്ധ്യത കൂടുതലാണോ ?[Are older people more likely to get coronavirus infection ?]	50 ഉം അതിൽക്കൂടുതലും പ്രായമായവർക്ക് മറ്റുള്ളവരെ അപേക്ഷിച്ച് അങ്ങേയറ്റം രോഗാതുരമായ അവസ്ഥയിലെത്തിച്ചേരാൻ രണ്ടരയിരട്ടി സാധ്യതയുണ്ട്.[People of age 50 and older are two and a half times more likely to become critically ill than others.]
4	കൊവിഡ്-19 ന് കാരണമാകുന്ന വൈറസ് വായുവിലൂടെ പകരുമോ?[Can the virus that causes covid-19 be spread through the air?]	പഠനങ്ങൾ സൂചിപ്പിക്കുന്നത് കൊവിഡ്-19 ന് കാരണമാകുന്ന വൈറസ് പ്രധാനമായും പകരുന്ന വായുവിലൂടെയല്ലാതെ ശ്വസന തുള്ളികളുമായുള്ള സമ്പർക്കത്തിലൂടെയാണ് എന്നാണ്. **[Studies indicate that the virus that causes covid-19 is mainly transmitted through contact with respiratory droplets rather than air.]**

Table 1: Sample Queries and Retrieved Results

The evaluation of the system is performed by using F1 Score. The F1 Score of the system is evaluated as 76 %.

6. Conclusion and Direction for Future Research

In the context of addressing the unique information needs of the native population during the ongoing COVID-19 pandemic, our work introduces a groundbreaking solution: a Semantic Malayalam Question-Answering system specifically tailored for COVID-19 care. The primary objective is to facilitate easy access to accurate and reliable information for Malayalam-speaking individuals, thereby contributing to their well-being during these challenging times.

Key features of our proposed system include the utilization of semantic mapping and modelling techniques for documents. We employ a Word Embedding model, specifically the Continuous Bag of Words (CBOW) approach, to achieve semantic representations of Malayalam text. This enhances the system's ability to comprehend the nuances of the language and extract meaningful relationships between words, crucial for accurate information retrieval.

The core mechanism for answering queries involves the application of Cosine Similarity. This metric is employed to assess the similarity between the semantic representations of the query and relevant documents in the dataset. By employing such advanced techniques, our system ensures a nuanced and accurate matching process, ultimately leading to more precise and contextually relevant responses. To evaluate the system's efficacy, we conducted experiments using our proprietary Malayalam question-answering dataset, meticulously curated from reputable sources. The results, as evidenced by the F1 score of 76%, attest to the system's reliability and performance in providing trustworthy information related to COVID-19 in the Malayalam language.

As part of our future direction, we are committed to enhancing the effectiveness of our Malayalam word embedding model. This involves exploring and implementing more advanced methods to further refine the semantic representation of the language. This ongoing improvement aims to bolster the system's understanding of the Malayalam language intricacies, ensuring its adaptability to a wider range of queries.Additionally, our future work extends beyond the confines of Malayalam and into the realm of open domain questioning. We aspire to broaden the scope of the system by experimenting with diverse domains and potentially expanding its language capabilities to cater to a more extensive linguistic audience.

In summary, our Semantic Malayalam Question-Answering system not only addresses the immediate information needs of the native population concerning COVID-19 care but also lays the groundwork for continuous improvement and expansion into broader linguistic and domain contexts in the future.

References

1. Green BF, Wolf AK, Chomsky C, and Laughery K. “Baseball: An automatic question answerer”. In Proceedings of Western Computing Conference, Vol. 19, 1961, pp. 219–224. Proceedings of AFIPS Conference, Vol.42, 1973, pp. 441–450.
2. YuanZhang, Dong Wang, Yan Zang, “Neural IRM eets Graph Embedding: A Ranking Model for Product Search” The Web Conference, May 2019, USA, ACM.
3. Piyush Mital, Saurabh Agrawal, Bhargavi Neti, Yashodhara Haribhakta, Vibhavari Kamble, Krishnanjan Bhattacharjee, Debashri Das, Swati Mehta, Ajai Kumar. “Graph-based Question Answering System”, ICACCI Dec 2018 IEEE
4. Tom Young , Devamanyu Hazarika , Soujanya Poria , Erik Cambria. “Recent Trends in Deep Learning Based Natural Language Processing”, arXiv, Nov 2018, August 2018, IEEE Computational intelligence magazine.
5. Fan fang, Bo-wen zhang, and Xu-cheng yin. “Semantic Sequential Query Expansion for Bio- medical Article Search”, 2169-3536 2018 IEEE.
6. Bo Xu, Hongfei Lin, Yuan Lin. “Learning to Refine Expansion Terms for Bio-medical Information Retrieval Using Semantic Resources” , 10.1109/TCBB.2018.2801303, IEEE/ACM.
7. Xu, B., Lin, H., Lin, Y. (2016). “Assessment of learning to rank methods for query expansion”. Journal of the Association for Information Science and Technology, 2016, 67(6): 1345- 1357.
8. DwaipayanRoy, Debasis Ganguly ,sumit Bhatia ,Srikanta Bedathur,Mandar Mitra, “Using Word Embeddings for Information Retrieval: How Collection and Term Normalization Choices Affect Performance”, 3269206.3269277 CIKM ’18, October 22–26, 2018, Torino, Italy, ACM.

9. Shomi Khan , Khadiza Tul Kubra, Md Mahadi Hasan Nahid, “Improving Answer Extraction For Bangali Q/A System Using Anaphora-Cataphora Resolution”, International Conference on Innovation in Engineering and Technology (ICIET) 27-29 December, 2018 IEEE.
10. Archana S.M. , Naima Vahab , Rekha Thankappa , C. Raseek,”A Rule Based Question Answering System in Malayalam corpus using Vibhakthi and POS Tag Analysis”,International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST – 2015) .
11. Arjun Babu, Sindhu L. “An Information Retrieval System for Malayalam Using Query Expansion Technique”,978-1-4799-8792-4/15/$31.00 c 2015 IEEE
12. Vaishali Singh,Sanjay K. Dwivedi .”Personalized approach for automated question answering in restricted domain”.International Journal of Information Technology, Springer.https://doi.org/10.1007/s41870-018-0200-6. 2018.
13. Sheetal S. Sonawane , Parag Kulkarni. Concept based document similarity using graph model International Journal of Information Technology,Springer. https://doi.org/10.1007/s41870-019-00314-w. 2019.
14. Swathilakshmi Venkatachalam , Lakshmana Pandian Subbiah et al[3].”An ontologybased information extraction and summarization of multiple news articles”. International Journal of Information Technology,Springer. 2019.
15. Navjot Kaur, Himanshu Aggarwal.”Query reformulation approach using domain specific ontology for semantic information retrieval”. International Journal of Information Technology, Springer. https://doi.org/10.1007/s41870-020-00464-2. 2020.
16. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi “Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis”, JBHI.2017.2767063,IEEE.
17. T. Kawamura, K. Kozaki, T. Kushida, K. Watanabe, and K. Matsumura, ‘‘Expanding science and technology thesauri from bibliographic datasets using word embedding, ’’in Proc. IEEE Int. Conf.Tools Artif. Intell., Nov. 2017, pp. 857–864
18. Liji S K and Lajish V L., “An Efficient Malayalam Query Processing System for University Enquiry “, Proceedings of the Eight National Conference on Indian Language Computing (NCILC), 2018, March 2018, CUSAT, Kerala.
19. Roberto Passailaigue Baquerizo ,Hubert Viltres Sala , Paúl Rodríguez Leyva , Vivian Estrada. “Sentí :Model for semantic processing in information retrieval systems”, International Research Journal of Engineering and Technology (IRJET),Volume: 04 Issue: 05 ,May -2017.
20. Liji S K, Muhamed Ilyas P, "Review and Analysis of Different Approaches to Semantic Level Question Answering and Information Retrieval", International Journal of Science and Research (IJSR),https://www.ijsr.net/search_index_results_paperid.php?id=SR21121141135, Volume 10 Issue 1, January 2021, 1238 – 1244.