

Answer Retrieval System using NLP for Documents: A Promising Solution for Research Queries

Prashant Y Niranjani¹, Vijay S. Rajpurohit², Sanjeev S Sannakki³, Kuldeep Sambrekar⁴, Ramesh Medar⁵, Girish R Deshapande⁶ and Prakash Sonawalkar⁷

¹Department of Computer Science & Engineering, KLS Gogte Institute of Technology, Belagavi
Email: prashant053@gmail.com

²⁻⁷Department of Computer Science & Engineering, KLS Gogte Institute of Technology, Belagavi
Email: {vijaysr2k@yahoo.com, sannakkisanjeev@yahoo.co.in, kuldeep.git@gmail.com, Rameshmedar2184@gmail.com, grdeshpande@git.edu, profsonwalkar@gmail.com}

Abstract— This paper presents a study of answer retrieval systems, which aim to automatically retrieve the most relevant answer(s) to a given question from a large collection of unstructured text data. Answer retrieval is a critical component in question-answering systems and has wide-ranging applications, from search engines to chatbots. We begin by providing an overview of the various techniques used in answer retrieval systems, including information retrieval models, neural network-based methods, and hybrid approaches. We then discuss the specific challenges faced by answer retrieval, such as the need to handle complex queries, the impact of noise and redundancy in the data, and the trade-offs between precision and recall. Next, we review the major datasets and evaluation metrics used in the field, highlighting their strengths and weaknesses. Finally, we discuss recent advances in answer retrieval systems, including the use of pre-trained language models, domain-specific knowledge, and multi-stage retrieval pipelines. Through our analysis, we identify key research directions for the continued development of answer retrieval systems, including the integration of conversational context, the use of multi-modal data, and the development of more effective evaluation metrics.

Index Terms— Answer Retrieval System(ARS), Bidirectional Encoder Representations from Transformers(BERT), Jensen-Shannon(JS), Document Similarity(DS), Stop Words, Gensim, Corpus, Best Matches 25(BM25).

I. INTRODUCTION

Answer Retrieval System is a combination of Question Answers and Information Retrieval, compliant users to get correct answers and admissible info about their query. In the search engine, the query is conferred in the form of a few keywords and a set of links or info that will pick up the answer aback to the user. Users must have spend a lot of time investigate this info and links to get the legitimate answer. Whereas, in an answer retrieval system, users demand questions in definitive languages and then assign the accurate answer to the user. CARA and Dolphin are examples of surveys that mainly focus on closed sites. Some popular examples of QAS are ChatGPT, Amazon's Alexa, Google Search, etc. ARS need machine learning, info renewal and natural language processing. Questions from users are generally branched into appropriate query and inaccurate questions. Factoid questions return a feedback this is a basis or noun phrase. Non-factoid questions generally prop descriptive answers to questions.

Closed domain question answering systems are a type of artificial intelligence technology that aims to automatically answer questions in a specific domain or area of knowledge. These systems are designed to analyse a given corpus of text or data, and then use natural language processing (NLP) techniques to understand the meaning of questions asked by users, and provide precise and relevant answers.

In a closed domain QA system, the input question is analysed and compared with the available knowledge base or dataset to find the most relevant answer. This system is different from an open domain QA system, where the system tries to find the answer from a wide range of knowledge sources, which can make the answer less accurate. In the context of research papers, closed domain question answering systems can be used to automatically retrieve information from a large corpus of scientific literature, and provide answers to specific research questions. The development of closed domain question answering systems for research papers can help researchers save time and effort by automating the process of information retrieval and analysis. Additionally, these systems can help researchers to discover new insights and connections in the field, by providing them with quick and accurate answers to complex research questions.

II. LITERATURE SURVEY

In community feedback systems, where we learn to use info possessed to use answer we can use theoretical approach to use acknowledgment, but ciphering out query and where they prowess detail to different answers and illuminating relevant answers can be gradual embracing to go through numerous answers to find what they're stretched toward for. To overcome this dilemma the following files are provided:

Author	Title	Year	Proposed Technique
Tripti Dodiya, Sonal Jain et al.	Question Classification for Medical Domain Question Answering System	2016	This article focuses on the adoption of the bi-level taxonomy [4] of 6 collections of granularity and 50 granular bands fabricated by Li and Roth, where we break down the questions into distinct categories.
A. Chandurkar and A. Bansal	Information Retrieval from a Structured KnowledgeBase	2017	This article intends to make question feedback more challenging than simple information retrieval[8] by search engines by approbation the confrontation of illuminating and authenticating accurate feedback.
Pradeep Kumar Roy, Zishan Ahmad, Jyoti Prakash Singh	Finding and Ranking High-Quality Answers in Community Question Answering Sites	2017	The article focuses on assorted aspects of feedback and end user reputation to predict relevant feedback Results were validated with absolute values for efficiency regression, field under the receiver operating characteristic curve, and root mean square error[6].
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova	Bert: Pre-training of deep bidirectional transformers for language understanding	2018	BERT is a simple and empirically flawless concept. It accomplished avant-garde results on 11 language processing tasks, including developing GLUE scores to 80.5% and MultiNLI [1] accuracy to 86%.
Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits	Is bert really robust? natural language attack on text classification and entailment	2019	This unit uses a ternary approach (i) to out-compete attainment conditions and demand impact;(ii) it preserves semantic content, grammaticality[2], and correct types classified by People; (iii) to Establish a favorable reaction loop alongside complication scales linearly including the magnitude of the text.

III. PROPOSED SYSTEM

The major conception of the system is to allow the most adequate and best feedback to the knowledge based on the current position. The generic architecture of the proposed system is shown in Figure 1.

The architecture consists of

1. **Document Store:** This System uses a document store to store and index documents. The document store is responsible for retrieving documents based on search queries.
2. **Pre-processing:** Before the documents can be indexed, they need to be pre-processed. This typically involves cleaning the text, removing stop words, filtering of words and tokenizing the text into individual words.
3. **Indexing:** This system uses the document store to index the documents. The indexing process involves creating an index that contains the document IDs and their corresponding vectors. The vectors are used to perform similarity searches to retrieve documents that are relevant to a given query.

4. *Retrieval*: When a user submits a search query, it retrieves the relevant documents from the document store using the index.
5. *Ranking*: The retrieved documents are then ranked based on relevance to the search query. This is typically done using a machine learning technique named Jensen-Shannon trained on a large corpus of documents.
6. *Answer Extraction*: The answer extraction component of this system uses a machine learning model to extract answers from the retrieved documents.
7. *Answer Formatting*: Finally, the extracted answers are formatted and returned to the user in a human-readable format.

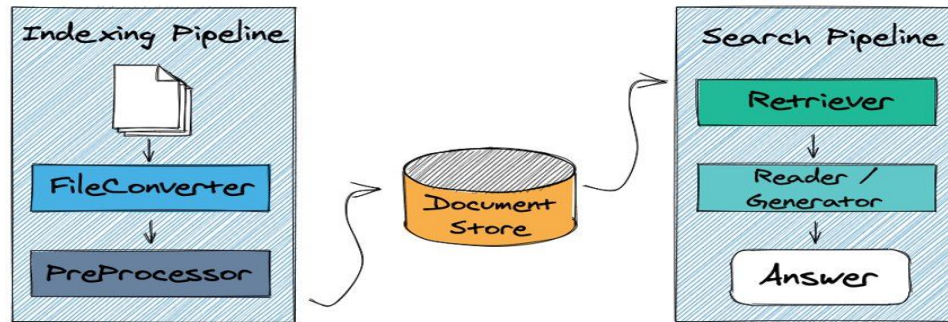


Figure 1. Block diagram for Answer Retrieval System

The sub-components of our proposed system are:

A. Domain Resources

The Domain Resources refer to admissible info and databases. These act as backends for the QA system. These resources are not directly accessed by user queries, but are accessed indirectly as a result of the Application Corpus is created by them.

It should be generic for where QA systems are installed. Because even if the info changes, there is no need to renovate the ontology. Domain datasets are info that end user will to question It can be in any format. Requirements are accurate spelling, accurate grammar and well-structured sentences.

The two subcomponents mentioned above are detached and establish in the info document each word is remarkable with an ontology site ID. There will be multiple matches or no matches at all. This file consists of the application corpus in the corpus component of the QA system.

B. Corpus

This is a titled document constructed from a domain resource. Questions from end user are formulated into a corpus [10] after passing through the classification phase. This is done with the help of a query-corpus mapper. The query-corpus [15] mapper breaks down the query and extracts the features. These features are then in comparison contrary to the same criteria extracted from each sentence in the document. Mappers access these properties by interacting with the Application Assembly. Then find the closest match and flag it the ultimate frequent sentence.

IV. METHODOLOGY

The methodology for building a question answering system can vary depending on the specific application and the technologies used. Some systems may rely on deep learning models, such as transformers, to perform the question analysis, document retrieval, and answer extraction tasks. Others may use simpler rule-based or statistical techniques for these tasks. A question answering system typically follows a methodology that involves several steps. The first step is to pre-process the input data, which involves cleaning and tokenizing the text. Once the question is classified, the system retrieves relevant documents from a pre-defined corpus of data. The system then extracts the most likely answer to the question using various techniques, such as named entity recognition or machine learning models like BERT. After an answer is extracted, the system checks whether it is valid and accurate. Finally, the system presents the answer to the user in a human-readable format, such as a sentence or a paragraph. The specific methodology used may vary depending on the application and the type of questions being asked.

The features of the question are matched with features of various parts of the document. This step filters out the irrelevant parts. Thus, most relevant answers are returned to the end user.

Some of the methods which we used are as follows:

A. Bidirectional Encoder Representations from Transformers(BERT)

It is a pre-trained transformer-based language model developed by Google. BERT is designed to understand the context of words in a sentence and is capable of producing high-quality representations of text for various NLP tasks, including question answering, text classification, and named entity recognition.

Using BERT in a question commonly involves the following steps:

- i. **Text input pre-processing:** The data file is pre-processed by cleaning it and modify it to the form the BERT arose from.
- ii. **BERT fine-tuning on definite info:** BERT fine-tunes secure query-response info, the processing stage a custom rectification method is added on top of the pre-trained BERT model, which is trained to answer questions through feedback and generate the correct answer.
- iii. **Identify the priority of info documents:** When a user enters a question in the question field, BERT is used to identify the most relevant data files that may contain an answer. BERT can do this by understanding question content and text files.
- iv. **Extracting answers from the text:** After the gratified of the text is fixed BERT is used to extract the feedback from the text.
- v. **Response validation:** Finally, validate the feedback from BERT to ensure it is accurate and valid.

B. Document Similarity

Document similarity refers to the degree to which two or more documents are similar [11] or related to each other in terms of their content or meaning. It is a measure of how closely related the topics, themes, or ideas in the documents are.

Document similarity [9] can be useful in various applications, such as information retrieval, text mining, and natural language processing. For example, in information retrieval, document similarity can be used to retrieve relevant documents based on the similarity between the query and the documents in a corpus. The workflow of document similarity for our system is as shown in the figure given below:

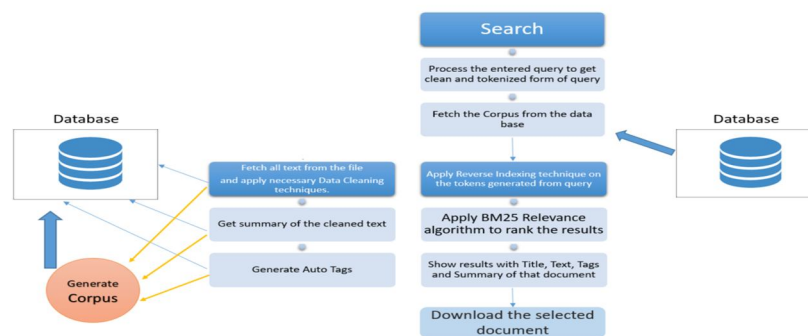


Figure 2. Workflow of Document Retrieval

C. Jensen-Shannon Technique

Jensen-Shannon approach The Jensen-Shannon approach also known as Jensen-Shannon deviation (JSD)[17], is a mathematical method used to measure the similarity between two proportion It is a lateral and smoothed version of the Kullback-Leibler (KL)[20] deviation, which is a measure of the difference between the two probabilities.

The JSD is calculated by first computing the average of two probability distributions and then calculating the KL divergence of each distribution from the average distribution.

The JSD is the square root of the average of the two KL divergences. The JSD ranges between 0 and 1, with a value of 0 indicating that the two distributions are identical, and a value of 1 indicating that the two distributions are completely dissimilar.

The Jensen-Shannon distance between the two probability vectors p and q is defined as Search by Key, Label, and Name.

$$\sqrt{\frac{D(p \parallel m) + D(q \parallel m)}{2}} \quad (1)$$

The JSD technique is commonly used in natural language processing (NLP) applications to measure the similarity between two sets of documents or texts. For example, in document clustering or text classification, JSD can be used to measure the similarity between the probability distributions of different documents or text categories. It can also be used in topic modelling to measure the similarity between different topics or clusters of topics.

D. Gensim Library

Gensim is a Python library that can be used for document retrieval. It first pre-processes the documents by removing noise such as stop words and punctuation [13]. Then it converts the pre-processed documents into numerical vectors using techniques such as Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF)[14]. Next, the vectorised documents are indexed using algorithms such as Latent Semantic Indexing (LSI) or Latent Dirichlet Allocation (LDA)[14]. When a user queries the system, the query document is pre-processed and vectorised in the same way as the indexed documents. The similarity between the query document and the indexed documents is calculated using techniques such as cosine similarity. Finally, the indexed documents are ranked based on their similarity to the query document, and the most similar documents are returned as search results. Gensim's efficient algorithms for pre-processing, vectorization, indexing, and similarity calculation make it a useful tool for document retrieval tasks.

For example:

You have a document which consists of words like - cup, car, keys, liquid, glass, racket, steering, score, bat, water. These can be grouped into different topics as-

Topic 1	Topic 2	Topic 3
glass	bat	car
cup	racquet	drive
water	score	keys
liquid	game	steering

V. FINDINGS AND ANALYSIS

Using the analytical text of 5000 info collected from Data, the accuracy [12], precision, and retreat of BERT for return feedback were calculated. These calculated values are shown below.

It will be to estimate Bidirectional Encoder Representations from Transformers (BERT) with other Machine Learning (ML) classification algorithms. These algorithms are ML classification algorithms that will be compared: DistilBERT, K-Nearest Neighbors (KNN), Multinomial Naive Bayes, Support Vector Model (SVM).

To compare these algorithms, we will:

- Feature Engineering: create features according to the raw data.
- Dissect and understand the data made available.
- Pre-process these data according to the algorithm: for case, some of these algorithms only work with numerical values.
- Fine-Tuning: optimize training parameters of the ML algorithm.
- Compare the results attained.

If we epitomize the results attained, here is what we get

We can see that BERT and other models are the ML classification algorithms that provide the best results. However, based on the scores, we can see that there is a significant difference in precision and accuracy between these algorithms.

ML Algo	Accuracy	Precision	Recall
BERT	99.910	100.000	99.329
DistilBERT	99.641	98.658	98.658
KNN	95.695	98.095	69.128
Multinomial Naive Bayes	98.117	95.070	90.604
SVM	98.655	98.551	91.275

Figure 6. Numerical representation of various parameters for different ML Models



Figure 7. Graphical representation of various parameters and comparison between different models

VII. CONCLUSION

Text allocation is a critical and imperative part of answer This article presents indefinite approach for classifying text. We evaluate the use of various algorithms, machine learning and neural network classifiers for arrangement dilemma and find that BERT gives excellent outcome for most of the text. Different models do different possessions for distinctive query A revolution that performs well in single corpus may not achieve as well in a different corpus.

These type of systems are an important application of natural language processing (NLP) that aim to automatically answer user's query's by retrieving relevant information from large text collections. These systems typically use machine learning models such as BERT and document similarity techniques such as Jensen-Shannon distance to match user's questions with relevant information from the text. They also use various techniques to pre-process the text, extract features, and rank the results, such as Named Entity Recognition (NER) and Passage Retrieval.

REFERENCES

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [2] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? natural language attack on text classification and entailment. arXiv preprint arXiv:1907.11932, 2019.
- [3] Tripti Dodiya, Sonal Jain et al., Question Classification for Medical Domain Question Answering System, 2016 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE) 19-21 December 2016, AISSMS, Pune, India.
- [4] Pradeep Kumar Roy, Zishan Ahmad, Jyoti Prakash Singh, "Finding and Ranking High-Quality Answers in Community Question Answering Sites", Global Journal of Flexible Systems Management Springer, November 2017.
- [5] A. Chandurkar and A. Bansal, "Information Retrieval from a Structured KnowledgeBase," 2017 IEEE 11th International Conference on Semantic Computing (ICSC), San Diego, CA, 2017, pp. 407-412.
- [6] Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, Georg Rehm, "Aspect-based Document Similarity for Research Papers," Proceedings of the 28th International Conference on Computational Linguistics, pages 6194–6206 Barcelona, Spain (Online), December 8-13, 2020.
- [7] ELAINE VAUGHAN and ANNE O'Keeffe, "Corpus Analysis" The International Encyclopedia of Language and Social Interaction, John Wiley & Sons, 2015.

- [8] Lee, Michael D, Pincombe, Brandon Welsh, Matthew, "An Empirical Evaluation of Models of Text Document Similarity " Proceedings of the Annual Meeting of the Cognitive Science Society, 27(27), 2005.
- [9] Izzeddin Gur, Daniel Hewlett, Alexandre Lacoste, Llion Jones, "Accurate Supervised and Semi-Supervised Machine Reading for Long Documents", Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages, 2011–2020, 2017.
- [10] Charles Yu, Ryan Sie, Nico Tedeschi, and Leon Bergen, "Word Frequency Does Not Predict Grammatical Knowledge in Language Models" Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4040–4054, 2020.
- [11] Juee Gosavi and Prof. B.N.Jagdale, "Answer Selection In Community Question Answering Portals" International Journal of Computer Engineering and Applications, Volume XII, Special Issue, May 18, www.ijcea.com ISSN 2321-3469, 2018.
- [12] J.-D. Kim , T. Ohta, Y. Tateisi and J. Tsujii, "GENIA corpus—a semantically annotated corpus for bio-textmining", Bioinformatics 19(1) c Oxford University Press 2003.
- [13] Ali Mohamed Nabil Allam and Mohamed Hassan Haggag, "The Question Answering Systems: A Survey", International Journal of Research and Reviews in Information Sciences (IJRRIS) Vol. 2, No. 3, September 2012.
- [14] Luciano Zunino, Felipe Olivares, Haroldo V.Ribeiro, and Osvaldo A.Rosso "Permutation Jensen-Shannon distance: A versatile and fast symbolic tool for complex time series analysis", arXiv:2203.12107v1 23 Mar 2022.
- [15] Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF, "DistilBERT, a distilled version of BERT:smaller, faster, cheaper and lighter", arXiv:1910.01108v4 1 Mar 2020.
- [16] Hung Chim and Xiaotie Deng, "Efficient Phrase-Based Document Similarity for Clustering", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 9, SEPTEMBER 2008.
- [17] Richard Connor, Franco Alberto Cardillo, Robert Moss, and Fausto Rabitti, "Evaluation of Jensen-Shannon Distance over Sparse Data", N. Brisaboa, O. Pedreira, and P. Zezula (Eds.): SISAP 2013, LNCS 8199, pp. 163–168, 2013. c Springer-Verlag Berlin Heidelberg 2013.
- [18] Mofiz Mojib Haider, Md. Arman Hossin, Hasibur Rashid Mahi, Hossain Arif, "Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm", 2020 IEEE Region 10 Symposium (TENSYP), 5-7 June 2020, Dhaka, Bangladesh.