IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Question Answering System Approaches: A Review

Mandar Suryavanshi
Department of Computer Science, Mithibai College, Mumbai, Maharashtra, India
University of Mumbai University of Mumbai

Abstract: *Data is increasing in volume day by day, this data can be processed and classified into various categories. Users all over the internet ask tons of questions to which they want precise answers. The Question Answering system is the best solution in such scenarios. Traditional approach mainly focuses on providing the documents which include the keywords related to the query asked by the users. While the question answering approach provides a better alternative by further refining the results, it not only returns the related documents but also retrieves the most relevant answer from the available corpus of data. Converting the questions asked by the users into an appropriate query string, classifying the question, retrieving the documents and extracting the valid answer are the main steps involved in this system.*

Keywords: Question answering system, Natural language processing, Information retrieval, Question Classification, QA system Approaches

I. INTRODUCTION

Question answering system is a combination of various fields like Natural language processing, Information retrieval, Information extraction, etc. It can be used to rank the documents in the order of relevance and also extract multiple answers for a single query; these answers can be further ranked according to the context required by the user. Natural language processing is used to convert the question asked by humans in a query format, Information retrieval scans through the data corpus to find related documents and Information extraction is used to formulate the final answer from the shortlisted documents.

The main objective of the question answering system is to provide answers rather than just ranking the documents. The data used in such systems can be both structured or unstructured. It consists of two types of domains namely open domain and closed domain. Open domain can have data related to any topic available over the web while closed domain is restricted to a specific topic like geographical questions, weather forecasting, etc.

Systems available in the past were just capable of answering factoid questions, for example “What’s the capital of India?” This would return a single word answer i.e., “Delhi”. However modern systems are more capable than their previous versions. Now they can answer in paragraphs, based on the type of question asked by the user.

Most common steps in such systems include question parsing- classifying the type of question, document analysis- shortlisting the relevant documents, answer analysis- extracting the answer from the most document. There are various approaches which can be used in a question answering system namely Linguistic approach, Statistical approach and Pattern based approach which can further be divided into subcategories.

This paper discusses the general framework used in Question answering systems along with some major approaches and techniques. It also gives us an idea about the scenarios suitable for each of them.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

II. FRAMEWORK OF QAS

A question answering (QA) system typically consists of several main components:

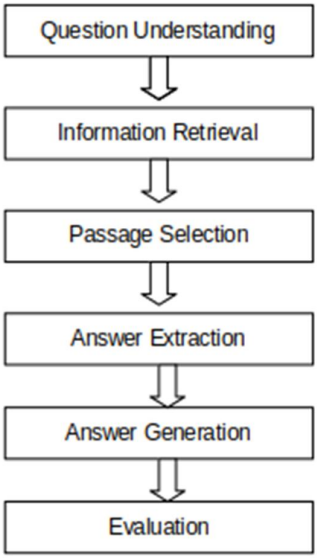

Fig.1.Components of QAS

2.1. Question Understanding: This component is responsible for understanding the natural language of the question and extracting the relevant information from it. This can be done using techniques such as natural language processing (NLP) and information extraction.

2.2. Information Retrieval: Once the question has been understood, the QA system uses this information to retrieve relevant documents or data from a pre-defined corpus or database. This can be done using techniques such as keyword search, Boolean retrieval, and vector space models.

2.3. Passage Selection: This component is responsible for selecting the most relevant passage(s) from the retrieved documents or data that contain the answer to the question. This can be done using techniques such as information retrieval, machine learning and deep learning models.

2.4. Answer Extraction: After selecting the relevant passage, the QA system uses this information to extract the answer to the question. This can be done using techniques such as information extraction, named entity recognition, and semantic role labeling.

2.5. Answer Generation: This component is responsible for generating the final answer to the question in a natural language format. This can be done using techniques such as natural language generation (NLG) and template-based generation.

2.6. Evaluation: The QA system's performance is evaluated by comparing the generated answer to the question with the ground truth answer using metrics such as precision, recall, and F1 score.

Overall, the QA system framework is designed to process natural language questions and extract the relevant information from the available dataset to give the most accurate response

III. ISSUES OF QAS FRAMEWORK

There are several issues that can arise in the framework of a question answering (QA) system, including:

3.1. Ambiguity: Natural language questions can be ambiguous and may have multiple interpretations, making it difficult for the QA system to understand the intent of the question and extract the correct information.

3.2. Vocabulary: The QA system may struggle to understand questions that use uncommon words or phrases, or that are expressed in a colloquial or informal style.

3.3. Lack of context: The QA system may not have enough context about the question or the dataset to generate an accurate answer.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

3.4. Complex or open-ended questions: The QA system may not be able to handle complex or open-ended questions that do not have a clear structure or set of possible answers.
3.5. **Limited dataset**: The QA system may not be able to answer a question if the information is not present in the dataset it has been trained on.
3.6. Biases: The QA system may be biased in its responses if the dataset used to train it contains biased information.
3.7. **Lack of evaluation**: The QA system may be trained and deployed without proper evaluation, causing suboptimal performance in real-world scenarios.
3.8. Lack of scalability: The QA system may not be able to handle a large amount of questions and data.
To overcome these issues, QA systems need to be developed with robust natural language processing (NLP) techniques, and be trained on large, diverse, and unbiased datasets. Additionally, the QA system should be continuously evaluated and fine-tuned to improve its performance and address any issues that arise.

IV. APPROACHES

4.1 Linguistic Approach

The linguistic approach to question answering involves using linguistic knowledge and techniques to understand and generate natural language. This approach aims to understand the meaning and structure of the question in order to provide an accurate and relevant answer.
One key aspect of the linguistic approach is the use of natural language understanding (NLU) techniques, such as syntactic and semantic analysis, to analyze the structure and meaning of the question. This allows the system to identify key elements of the question, such as the subject, predicate, and object, and to understand the intent behind the question.
Another important aspect is the use of natural language generation (NLG) techniques to generate an appropriate answer. This involves identifying relevant information in a knowledge base or other data source and using NLG techniques to generate a natural language response that is grammatically correct and semantically appropriate.
The linguistic approach also utilizes various NLP techniques such as Named Entity Recognition (NER), Part-of-Speech tagging (POS), Dependency Parsing, Coreference Resolution. These techniques are crucial to understand the context of the question and to generate an accurate answer. There are also various techniques like information retrieval and information extraction which are used to extract the information from the text corpus.
Overall, the linguistic approach to question answering systems is a powerful method for understanding and generating natural language, and can provide accurate and relevant answers to a wide range of questions.

Cons:
1) Complexity: Linguistic approaches can be complex and require a deep understanding of NLP and computational linguistics, making them difficult to implement and maintain.
2) Time-consuming: Linguistic approaches can be time-consuming, especially for complex questions that require a deep analysis of the language and context
3) Limited scalability: Linguistic approaches may not be scalable to large amounts of data and may become slow and inefficient as the size of the knowledge base grows.

QA system	Domain	Description
BASEBALL by Green et al [1]	Closed domain	Answering Question about Baseball game Front ends to databases
LUNAR Woods [2]	Closed Domain	Compare and evaluate the chemical analysis data on lunar rock and soil.
ELIZA Joseph Weizenbaum [3]	Closed Domain	Attempt to mimic basic human interaction Question and answer exchanges
GUS	Closed	A frame-driven dialog system

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Bobrow et al. [4]	Domain	Genial Under stander system also used structured database as the knowledge source
Clark et al [5]	Closed Domain	knowledge-base question answering ability through inference engine component
STARTQA System Boris Katz. [6]	Closed Domain	Web-Base QA system the system can answer millions of English questions about places (e.g., cities, countries, Etc.)
Mishra et al [7]	Closed Domain	Web documents in the local knowledge database
Quarc Rilloff et al [8]	Closed Domain	Rule-base QA system for reading comprehension tests

Table.1. Linguistic based QA systems

4.2 Statistical Approach

A statistical approach in a question answering system involves using machine learning algorithms and large amounts of data to train the system to understand natural language questions and provide accurate and relevant answers. It uses statistical models to determine the most likely answer. These models can be based on various techniques such as supervised learning, unsupervised learning, and deep learning. Commonly used algorithms for question answering systems include neural network-based models such as transformer, LSTM, and CNN. These systems can be trained on large datasets of questions and answers, and can be fine-tuned for specific domains or tasks. Maximum entropy model, Support vector machine classifier, Bayesian Classifiers are few common techniques applied by Statistical based QA systems.

Cons:

1) Data dependency: The performance of a statistical approach is heavily dependent on the quality and quantity of the training data used. If the training data is limited or of poor quality, the system may not be able to accurately answer questions.

2) Bias in the data: If the training data contains biases, the system may reflect these biases in its answers, leading to incorrect or unfair results.

3) Overfitting: Statistical models can overfit to the training data, leading to poor performance on new or unseen data.

QA system	Domain	Technique	Description
Ittycheriah et al [9]	Open	Maximum Entropy Model	Maximum entropy model for question/ answer classification based on various N-gram or bag of words features.
Cai et al [10]	Open	Sentence Similarity Model	Web-base Chinese QA system with answer validation
Soricut et al [11]	Open	N-gram mining	used a statistical chunker questions Into chunks/phrases asked to the search engine
Rocchio Moschitti [13]	Open	Support Vector Machine text classifier	Question and answer categorization and tested his approach on Reuters-21578.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Zhang et al [14]	Open	Support Vector Machine based on the features of words	Chinese QA system with question classification and answer clustering

Table. 2. Statistical approach based various QA systems and their technique

4.3 Pattern Matching Approach

A pattern-based approach in a question answering system involves using a set of predefined patterns or rules to identify the meaning of a question and extract relevant information to generate an answer. These patterns or rules can be based on various factors such as the structure of the question, the words used, and the context of the question. In a pattern-based approach, the system will have a set of predefined patterns for different types of questions, based on that it can answer the question. This approach is mainly used in the early stage of QA systems.

One of the advantages of a pattern-based approach is that it can be relatively simple to implement and can be fine-tuned for specific domains or tasks. However, the performance of a pattern-based system may be limited by the quality and coverage of the patterns or rules used. And this approach may not be able to handle out of domain questions or new types of questions.

Cons:

1) Limited coverage: Pattern-based approaches may not be able to handle a wide range of questions, especially if the set of patterns or rules is limited or not comprehensive enough.

2) Inflexibility: Pattern-based approaches can be rigid and inflexible, as they rely on predefined patterns or rules. They may not be able to handle new or unexpected types of questions.

3) Maintenance: Pattern-based approaches require regular maintenance and updating of the patterns or rules to keep up with changes in the language and new types of questions.

Pattern Matching approach	Surface Pattern Based	Finding answers to factual question answer are limited to one or two sentences
	Template based	Used for Closed domain focuses on interpretation

Table.3. Pattern Matching Approach

4.3.1 Surface Based Approach

A surface pattern-based approach in a question answering system is a specific type of pattern-based approach that focuses on identifying patterns in the surface level structure of the question, such as the words used and their arrangement. This approach typically involves using techniques such as natural language processing (NLP) to analyze the question and identify specific keywords or phrases that are indicative of the meaning of the question, and then using these keywords or phrases to extract relevant information from a knowledge base or other source of information. This approach relies on the surface level structure of the question, such as the keywords and grammar used to identify the intent of the question and then use that to extract the relevant information. This approach is mainly used where the system has a limited understanding of the language. Additionally, it may not be able to understand the context and the underlying meaning of the question

QA system	Descriptions
Hovy et al [15]	Learning surface text patterns for QA system Implemented an automatic learning environment
Soubbtin et al [16]	Pattern of potential answer expression as clues to the right answers.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Zhang et al [17]	Web-base pattern mining and matching approach to question answering
Greenwood et al [18]	Using name entity tagger to generalize surface matching text pattern for question answering
Cui at al [19]	Soft pattern matching model for definitional QA system. Bigram model and PHMM
Saxena et al [20]	Using pattern matching semantic type and semantic category. For difficult question

Table .4. Surface Pattern Based QA system

4.3.2 Template Based Approach

A template-based approach in a question answering system involves using pre-defined templates to generate responses to specific types of questions. The system is trained on a set of question-answer pairs that follow a certain template, and when it receives a new question, it tries to match the question to one of the templates it has been trained on. If a match is found, the system generates a response using the template, and if not, it may either generate a default response or flag the question as unanswerable. This approach can be effective for certain types of questions, such as those that have a clear structure and a limited set of possible answers, but may not be suitable for more open-ended or complex questions.

QA system	Technique	Description
Sneiders [21]	Frequently answer question (FAQ)	Automated QA using question template that cover the conceptual model of the database
Gunawerdena et [22]	Pre-processed text to identify best matched template-answer	A closed domain system to understand SMS language
Unger et al [23]	Resource description framework (RDF)	Template based QA system over RDF
SPARQL [24]	Resource description framework (RDF)	SPARQL query language for RDF

Table 5. Template based QA systems

4.3 Hybrid Approach

A hybrid approach in question answering systems combines multiple methods and techniques to provide the best possible answer to a question. The idea behind a hybrid approach is to leverage the strengths of different methods and overcome their limitations by combining them in a complementary way.

For example, a hybrid approach might use a pattern-based approach to quickly match the question to a relevant answer, and then use a linguistic approach to verify and fine-tune the answer based on the meaning of the question. Another example might be to use a statistical approach to identify the most relevant answer from a large corpus of data, and then use a pattern-based approach to verify and validate the answer.

The key advantage of a hybrid approach is that it can provide more accurate and relevant answers than any single method alone. By combining the strengths of different methods, a hybrid approach can handle a wider range of questions and provide answers that are more accurate and context-sensitive.

QA system	Domain	Technique	Descriptions

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Kwork et al MULDER [25]	Open	Based on integration of linguistic and statistical approach	Fully Automated General purpose QAS
Chakrabarti et al [26]	Open	Linguistic and pattern based	WorldNet structure to determine the answer type
Xia et al [27]	Closed	Rule-based & SVM classifier	An integrated approach for question classification in Chinese cuisine QA system
Lee YH et al ASQA [28]	Open	Surface pattern & entropy method	Complex question answering with ASQA. Deal with Definition and relation questions
Ferrucci D et al IBM's WATSON [29]	Open	Surface pattern & entropy	An overview of Deep QA Project

Table .6. QA System based on Hybrid Approach

Approaches [30]	Questions types	Domain
Linguistics approach	Factoid	Closed
Statistical approach	Complex non-factoid	Open
Pattern Approach	Factoid, definition, acronym, date of birth	Closed

Table.7. Comparison of Approaches

V. RELATED WORKS

ChatGPT is primarily based on the language modeling approach to question answering. This means that the model is trained to predict the next word in a sequence of text given the previous words. During inference, the model generates a response to a question by predicting the next word in the sequence given the question as input.

In this approach, the model generates a response by sampling from the learned distribution of words in the training data. The model has been trained on a large amount of text data, which allows it to generate responses that are contextually relevant and consistent with the information it has been trained on.

However, ChatGPT can also be fine-tuned on specific domains or topics to improve its performance on questions related to that domain or topic. This fine-tuning process can involve using other techniques, such as pattern-based matching or statistical analysis, to provide additional information to the model and improve its performance on specific types of questions.

Overall, ChatGPT is a powerful tool for building a question answering system and its language modeling approach provides a strong foundation for generating human-like text responses to a wide range of questions.

VI. CONCLUSION

This paper discussed the approaches and techniques related to question and answering systems. Various challenges and limitations still exist, in order to achieve better systems which can provide us with more accurate results compared to the current systems. Future work can be carried out to understand the subject deeper and propose new methods for enhancing the precision.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

ACKNOWLEDGEMENT

Thanks to the Computer Science Department of SVKM's Mithibai College, University of Mumbai. Thankful to the entire staff for providing support and facility for carrying out the research.

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.

[2]. Woods W. Progress in Natural Language Understanding - An Application to Lunar Geology. In Proceedings of AFIPS Conference, Vol. 42, 1973, pp. 441–450.

[3]. Weizenbaum J. ELIZA - a computer program for the study of natural language communication between man and machine. In Communications of the ACM, Vol. 9(1),1966, pp. 36-45.

[4]. Bobrow DG, Kaplan RM, Kay M, Norman DA, Thompson H, and Winograd T. Gus, a frame-driven dialog system. Artificial Intelligence, Vol. 8(2), 1977,pp. 155-173.

[5]. Clark P, Thompson J, and Porter B. A knowledge-based approach to question answering. In Proceedings of AAAI'99 Fall Symposium on Question-Answering Systems, 1999, pp. 43-51.

[6]. http://start.csail.mit.edu/index.php

[7]. Mishra A, Mishra N, Agrawal A. Context-aware restricted geographical domain question answering system. In Proceedings of IEEE International Conference on Computational Intelligence and Communication Networks (CICN), 2010, pp. 548-553.

[8]. Riloff E and Thelen M. A Rule-based Question Answering System for Reading Comprehension Tests. In ANLP /NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, Vol. 6, 2000, pp. 13-19.

[9]. Ittycheriah A, Franz M, Zhu WJ, Ratnaparkhi A and Mammone RJ. IBM's statistical question answering system. In Proceedings of the Text Retrieval Conference TREC-9, 2000.

[10]. Cai D, Dong Y, Lv D, Zhang G, Miao X. A Web-based Chinese question answering with answer validation. In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 499-502, 2005.

[11]. Soricut R and Brill E. Automatic question answering using the web: Beyond the factoid. In Journal of Information Retrieval-Special Issue on Web Information Retrieval, Vol. 9(2), 2006, pp. 191-206.

[12]. Suzuki J, Sasaki Y, Maeda E. SVM answer selection for open domain question answering. In Proceedings of 19th International Conference on Computational linguistics,COLING'02, Vol. 1, 2002, pp. 1-7.

[13]. Moschitti A. Answer filtering via text categorization in question answering systems. In Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, 2003, pp. 241-248.

[14]. Zhang K, Zhao J. A Chinese question answering system with question classification and answer clustering. In Proceedings of IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol.6,2010, pp. 2692-2696.

[15]. Ravichandran D and Hovy E. Learning surface text patterns for a question answering system. In proceeding of 40th Annual Meeting on Association of Computational Linguistics, 2002, pp. 41-47.

[16]. Soubbotin MM and Soubbotin SM. Patterns of Potential Answer Expressions as Clues to the Right Answer. In Proceeding of the TREC-10, NIST, 2001, pp. 175-182.

[17]. Zhang D and Lee WS. Web based pattern mining and matching approach to question answering. In Proceedings of the 11th Text Retrieval Conference,2002.

[18]. Greenwood M. and Gaizauskas R. Using a Named Entity Tagger to Generalise Surface Matching Text Patterns for Question Answering. In Proceedings of the Workshop on Natural Language Processing for Question Answering(EACL03), 2003, pp. 29-34.

[19]. Cui H, Kan MY and Chua TS. Soft pattern matching models for definitional question answering. In ACM Transactions on Information Systems (TOIS), Vol. 25(2):8, 2007.

[20]. Saxena AK, Sambhu GV, Kaushik S, and Subramaniam LV. Iitd-ibmirl system for question answering using pattern matching, semantic type and semantic category recognition. In Proceedings of the TREC, Vol.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

2007,2007.

[21]. Sneiders E. Automated question answering using question templates that cover the conceptual model of the database. In Natural Language Processing and Information Systems, Springer Berlin Heidelberg, 2002,pp. 235-239.

[22]. Gunawardena T, Lokuhetti M, Pathirana N, Ragel R,Deegalla S. An automatic answering system with template matching for natural language questions. In Proceedings of 5th IEEE International Conference on Information and Automation for Sustainability (ICIAFs),2010, pp. 353-358.

[23]. Unger C, Bühmann L, Lehmann J, Ngonga Ngomo AC, Gerber D and Cimiano P. Template-based question answering over RDF data. In Proceedings of the ACM 21st international conference on World Wide Web, 2012,pp. 639-648.

[24]. Prud'hommeaux E, Seaborne A(eds.). SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/, 2007.

[25]. Kwok C, Etzioni O and Weld DS. Scaling question answering to the Web. ACM Transactions on Information Systems (TOIS), Vol.19 (3), 2001, pp. 242-262.

[26]. Ramakrishnan G, Chakrabarti S, Paranjpe Dand Bhattacharya P. Is question answering an acquired skill?. In Proceedings of the 13th ACM international conference on World Wide Web, 2004, pp. 111-120.

[27]. Xia L, Teng Z, and Ren F. An Integrated Approach for Question Classification in Chinese Cuisine Question Answering System. In IEEE second International Symposium on Universal Communication, 2008, pp.317-321.

[28]. Lee YH, Lee CW, Sung CL, Tzou MT, Wang CC, Liu SH, Shih CW, Yang PY and Hsu WL. Complex question answering with ASQA at NTCIR-7 ACLIA. In Proceedings of NTCIR-7 Workshop Meetings, Entropy,1, 10, 2008.

[29]. Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, Lally A et al. Building Watson: An overview of the DeepQA project. AI magazine 31, no. 3,2010, pp. 59-79.

[30]. M., Ajitkumar, Khillare S.A., and C. Namrata. "Question Answering System, Approaches and Techniques: A Review." International Journal of Computer Applications 141.3 (2016): 34–39. Web.