Journal of Information Systems Engineering and Management
2024, 9(4s)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Bridging the Gap Between Question and Answer: A Comprehensive Analysis on Medical Question Answering Systems

Keerthana SM[1], Dr.K.VijayaKumar[2,3] ,Dr Mohammad Nazmul Hasan Maziz[4]
1 Assistant Professor,Department of Computer Science and Engineering ,St.Joseph's Institute of Technology, keerthanasmk@gmail.com
2 Professor,Department of Information Technology,St.Joseph's Institute of Technology,vijay@stjosephstechnology.ac.in
3 Postdoctoral Fellow, Faculty of Medicine and Health sciences, Perdana University, Malaysia
4 Professor,Faculty of Medicine and Health sciences, Perdana University, Kuala Lumpur , mohammad.nazmul@perdanauniversity.edu.my

ARTICLE INFO

Received: 23 Oct 2024

Revised: 15 Nov 2024

Accepted: 16 Dec 2024

ABSTRACT

Data is the universal language of information in real world. But according to a statistic only 20% of the data in real world are structured where remaining 80% of data are unstructured. The machine provides in accurate result in retrieving the information from these unstructured data. With help of Natural Language Processing (NLP), the machines are able to process the dark data and accomplish the task given by user. Among many tasks of NLP, Question Answering System (QAS)plays a vital role for the real-world development. QAS is the task of giving accurate answer for the question posted by user in natural language about the document (Textual Question Answering) or query about the image (Visual Question Answering) or question related to medical field (Medical Question Answering). This paper provides an overview of Medical QAS Datasets, Methodology implemented and the Metrics to evaluate the model. At the end of the survey, this paper provides a finalized overview of what methodology/approach can be used for the QAS.

Keywords: NLP,Question Answering System,Medical,Deep Learning

INTRODUCTION

Question Answering System is one of the NLP tasks, where a user post query or question in natural language and the system gives accurate answer for the posted question. In traditional method, NLP analyses the question and then retrieves answer using the information retrieval techniques, where a set of documents are listed instead of direct answer. This system became difficult and time consuming for the user to find the appropriate answer. The QAS are broadly divided into two types:1-open domain QAS, where the question can be about any field or domain and Large Language Models are used to generate the answer.2-Closed domain question answering, the question are being posted about a specific domain and answer is generated about the same. Apart from this there are four variants of QAS based on the input and output1-Factoid QAS,the questions are based on real world facts.Ex:How many bones does human have?, In which direction does the sun rise? 2-MCQ question answering, the questions are posted along with four options and the model has to select one correct answer from the option which is address as classification problem.3-Generative Question answering, given a text and question the model will generate the answer not based on given context.4-Extractive QAS,the model will extract the given text and produce the answer.In this paper we will discuss about the various methodology and approaches to optimize the models accuracy is discussed in Section I,the different Dataset in discussed in Section II that is used to train the model for visual QAS ,Medical QAS.Even though there is no specific metric for QAS ,precision ,recall,accuracy , F1 score has been used in various paper to evaluate the model,which is discussed in Section III.

METHODS

(Basu et al., 2020) says apart from retrieving the answer for medical question, understanding the semantics of the text is very important in the medical field. Given question into an ASP(Answer Set Programming) query using the framework[1] and then run it through the CASP goal-directed A00SP system. An ASP query is a request for

information or a solution formulated within the ASP framework. ASP is a declarative programming paradigm used for knowledge representation and problem-solving. In ASP, queries are typically expressed in logical terms, and the system attempts to find answer sets that satisfy the given query, based on the provided knowledge base and rules. These answer sets represent possible solutions or interpretations of the query within the defined constraints. VerbNet a novel algorithm based on partial tree matching generates an answer set program that represents the knowledge in the text.

Creating a open source multilingual model[2] that can support different language for diverse use of different communities. Firstly, they compile a multilingual medical corpus tailored for auto-regressive training, with the aim of establishing a sturdy base that faithfully captures the linguistic intricacies and diversity of the medical field. Secondly, to gauge progress, we introduce a comprehensive new multilingual medical question-answering (QA) benchmark. This facilitates evaluation of the multi-choice QA and rationale capabilities of various language models, both in zero-shot and fine-tuning scenarios. Lastly, we assess a broad range of existing language models, including those subjected to auto-regressive pre-training using our corpus.

As information retrieval method does not show great accuracy in retrieving the answer research has been carried out to retrieve answer from knowledge graph[3],to meet the high standard answer expectation of patient from doctor this model combines the knowledge graph and QAS.The user crawls through the vertical website and build a Knowledge graph consisting of disease and symptoms to retrieve the required answer grounded on rule based matching.As population is increasing the upgradation in medical field is also rapidly snowballing.The knowledge graph is created by the main object like symptoms and disease which is represented as entities and their relations are linked.The knowledge graph are built either using top down approach or bottom up approach.

Most of the QA community or forum we can see most of the complex real world questions are unanswered. The model is also not able to extract answer for the complex question which will remain ambiguous among user. Manually annotating unanswerable question is a time consuming process and hence a agent[4] is created to separately tag the unanswered question and then train the model to answer it.

In some cases the unanswered question are addressed through temporal knowledge[5], which adds timestamp to each of the entity .Each of the complex question is subdivided into simple question and answer is retrieved from the different knowledge graph. A pretrained language model (LM) to extract both the final hidden state and instruction vector. Subsequently, we integrate the temporal graph convolution network (T-GCN) module to extract implicit temporal characteristics from the semantic information encoded in the question representation.Most of the previous approaches have developed retrieval modules[6] for selecting relevant passages, they face challenges in scenarios beyond two hops.Limited Performance in one step hop and failure in two step hop of selecting irrelevant documents.Beam Retrieval, a general end-to-end retrieval framework for multihopQA

As there was lack in thorough understanding of the logical associations among the contents and structures of different documents. previous works train a multi-hop retriever[7,11] to imitate such process by sequentially fetching the next passage based on the already-retrieved ones, none of them explore the potential of engaging LLMs into this process.Proposal:Knowledge Graph Prompting (KGP) method to formulate the right context in prompting LLMs for MDQA, which consists of a graph construction module and a graph traversal module. The constructed graph serves as the global ruler that regulates the transitional space among passages and reduces retrieval latency. For the multi-modality challenge, we add different types of nodes.

Construction of and Usage in Question Answering for Clinical Practice Guidelines Problem:Clinical practitioner Guidelines CPG[8] is a document in medical filed for storing the information related to the history of disease and how it is cured.when a person suffers from prolonged Cancer and it has been cured by 50% at a stage, a document is made containing the information like what are the procedure followed during the treatment which will be useful for the doctors and practitioners to refer and give treatment in short time. Question answering on this documents does not

perform well because there is no huge data to train the model and when adaptations are made it is difficult to refer the upgradation manually.Proposed:Decision based knowledge graph helps in fast reference through the document and the if there is any upgradation it is easy to update the knowledge graph. Creation of dataset of triples containing 8300 questions from acute lymphoblastic leukemia, kidney, and bone cancer. Each triple consists of question, answer, and cypher query (used to query decision knowledge graph). The proposed model gives 40% better results compared to fine tuned transformer question-answering model.The answers are retrieved based on rule where a knowledge graph is kept as a high quality source ,the text from the knowledge graph is filled in the blank space based on the rule .This is achieved using GroupsteinerTress[9] a graph algorithm.

The Community platform has been serving as a information retrieval for Most researchers but sometimes the answer from ametur might mislead which will result in high medical issue or wrong medical guidance. In this paper REQUEST, a bi-directional autoregressive transformer[10] has been used.it first recognize the question by applying classification and then summarize the question based on thematic tags. The unanswered question can be clearly identified as the questions are tagged based on the topic The decoder part of BERT only first three layers are shared for the fine tuning of the model and the weighted loss are calculated.This method achieves a great result in generating answer for long,unambiquity question that are unanswered.

To retrieve question relevant evidences from RDF data fine tune BERT model has been used which works best for MCQ questions with factoid answers.This graph based method[11] provides user interpretable evidence for complex answering process.Before Knowledge graph patter based/rule based process was used. Retrieving answer based on patter alone is not enough due to emerging ephemeral facts**.** To enhance the answering capability of the model ,different knowledge base is grouped into single graph by entity alignment.If the knowledge base are unalignable entities may be treated as identical in question.Sometime due to incomplete knowledge base,the model might not retrieve correct answer which will lead ambiguity to the user.In specific the semantic parsing based methods face challenge like adaptability and interoperability.Based on the trained language model the reasoning path for answer is predicted.The entities that reasoning path pass through will reduce the distraction of the path.Later the path connecting to both the entities are forwarded to another model. Click or tap here to enter text.when the data was in structured query language it was easy to process but due to advancement in AI it is easier to process unstructured data. Neural machine translation runs out of vocalbury,which faces the new words that were not seen during training.

The entity linking is embedded from different knowledge source to reduce the model from getting biased or underfitted. During the entity linking phase[12] it create a query template to fill entities and slots are filled in which entity will fill the empty place. Information retrieval technique is used to extract information like entity and relation in the knowledge graph. The graph exploration navigates through entities of the knowledge base using rule based techniques.

A model identifies important medical content from patients record and then prepare few educational questions relevant to patient discharge summary. The dialogic reading of patient is done through PEER technique[13], in which it first Prompt and evaluate the input. The questions are expanded and repeated to extract correct answer. Ex.what was the cause of cardiac arrest? What type of cardiac catherization help treat a heart attack?

Struggling with multi-hop question answering over heterogeneous knowledge. Firstly, a complete multi-hop program relies on multiple heterogeneous[13] supporting facts, and it is difficult for models to receive these facts simultaneously. Secondly, these methods ignore the interaction information between the previous hop execution result and the current-hop program generation. Self-iterative framework for multi-hop program generation (HopPG) over heterogeneous knowledge, which leverages the previous-hop execution results to retrieve supporting facts and generate subsequent program. UniRPG, a SP-based model designed for HQA. e HopPG, an iterative program generation framework designed explicitly for multi-hop answer reasoning based on heterogeneous

knowledge.MMQA dataset for evaluating the effectiveness of HopPG. Specifically, we only focus on questions based on tables and texts, which we refer to as MMQA-T2.

Existing: LLM integration presents a challenge for small businesses due to the high expenses of LLM API usage. Costs rise rapidly when domain-specific data (context) is used alongside queries for accurate domain-specific LLM responses.Proposed: In this paper, we shift from humanoriented summarizers to AI model-friendly summaries. Our approach, Lean Context [14], efficiently extracts k key sentences from the context that are closely aligned with the query. The choice of k is neither static nor random; we introduce a reinforcement learning technique that dynamically determines k based on the query and context. The rest of the less important sentences are reduced using a free open source text reduction method.LLMs can learn domain-specific information in two ways, (a) via fine-tuning the model weights for the specific domain, (b) via prompting means users' can share the contents with the LLMs as input context.

Problem: Large Language Models can understand and communicate using language, promising richer human-AI interaction .Proposed: Med-PaLM 2[15], a new medical LLM trained using a new base model and targeted medical domain-specific finetuning. ensemble refinement as a new prompting strategy to improve LLM reasoning .Med-PaLM 2 achieved state-of-the-art results on several MultiMedQA benchmarks, including MedQA USMLE-style questions . Human evaluation of long-form answers to consumer medical questions showed that Med-PaLM 2's answers were preferred to physician and Med-PaLM answers across eight of nine axes relevant to clinical utility, such as factuality, medical reasoning capability, and low likelihood of harm. Two adversarial question datasets to probe the safety and limitations of these models. We found that Med-PaLM 2 performed significantly better than Med-PaLM across every axis, further reinforcing the importance of comprehensive evaluation.Answering Complex real world question are tedious compared to the general medical question or exams. Apart from textbook knowledge medical experts must be aware of real world scenarios to give moral support to the patients.Developing a model to give decision based on complex real life problem is a great challenge.Many research have been carried out to create a model to answer complex medical question posted by user in their own natural language. Most research paper uses Illama LLM[15] with different prompting approach where the first step involves in answering the question, and in second step the decision for the generated answer is prompted.Along with Prompting ,few shot learning is implemented so that new task can be learned without external training.

Table 1 Various Methodologies in Textual Question Answering System

Author	Dataset /Question Source	Methodology
(Basu et al., 2020)	SQUAD(factoid based question)	Generates explainable answers with semantic answer
(Qiu et al., 2024)	Medical Book of different languages(English,spanish,Chinese,Japanese,French, Russina,Spanish)	Monitor the development of MML
(Frisoni et al., 2024)	MedQA	Generates and then read the framework
(Jiao et al., 2023)	Bio Medical	Data augmentation to improve model accuracy and weighing strategies to extract the accurate answer
(J. Zhang et al., 2023)	HotpotQA & 2WikimultiHop	It supports for 2 Hops
(Y. Wang, Lipka, et al., 2023)	2wikiMQA,MusiQUE	A graph construction module & graph traversal graph
(Kandula & Bhattacharyya, 2023)	Dataset with triples	fast reference through document & update the graph
(Aftabi et al., 2024)	CQADup-Stack	Bi-directional auto aggressive transformer to recognize

		question,summarize and tag the question to fast answer retrieval
(Pramanik et al., n.d.-b)	RDF	Builds context graph to retrieve relevant evidences from RDF data
(M. Zhang et al., 2023)	MKBQA	
KBQA	Creates score & Rank candidate answer by creating link relation	Limitation of time & cost ,the dataset is limited to ~2.3k
(J. Zhang et al., 2023)	hopQA	Ene to end retrieval framework for multihopQA
(Cao et al., 2023)	WebQSP	Answers are retrieved using semantic parsing based & IR based method
(Diomedi & Hogan, 2021)	QALD/KGQA	Neural Machine Translation-to link the entity and form a a query template with place
(Cai et al., 2023a)	MIMIC III	Dialogic reading to help patient understand their discharge summary and gie feedback which focus on patient centric interative interative Question Answering
(Arefeen et al., 2023)	Arxiv & BBC	Extracts key sentences from context that are close to query
(Singhal et al., 2023)	MedQA	Combination of improved base LLM,medical domain specific fine tuning and novel prompting stratergy
(Chen et al., 2024a)	JAMA & MedBullets	Enhance high quality answer extraction from different sources
(Frisoni et al., 2024)	MedQA,MMLU,MedMcQA	First generate and then rea
(J. Wang et al., 2024)	MIMIC IV,AMBOSS	Reduces the demand for computational resources & enhances model ability for reasoning

MEDGENE[16] approach generates the answer for the question and then compares with the given option for the Multiple choice question.The answers are generated from Multiview background context to apply fusion in decode and fine tuning for LLM.When it comes to open domain QA ,the model must have access to extend the knowledge base to accomplish the task of answering the question.It learns only from the knowledge parameter and then move forward to learning from the knowledge base.The is input is augment with the relevant knowledge chuncks from external database like PUBMed and UMLS.Medical LLms have started to show increasing result to aid professional and improve health care.

As there is a increase in medical data and advancement in the field leads to more updation and accurate retrieval of the documents. They use RAG(retrieval Augmented Language Model)[17,25] to retrieve most accurate answer .As most LLM provide inaccurate answer ,they first train a retriever to get relevant document and then train LLM to generate a response. For information retrieval COLBERT is used which encodes both queries and documents. To retrieve answer from different documents, the document sequence is quoted and then queries shorter than predefined number of tokens are padded with BERT's special token to reach the length or else truncat.

Most of the LLM available today are proprietary model with hundred of billion of parameters that were trained on hundred of billion of tokens.LLM can accomplish the given task and show more accuracy by reducing the computation resource by PaLM[18] technique.The prompt is given as input to the LLM to get the output.The prompt is simple question which supports zero shot learning.To avoid the model from getting biased,each time the options

are shuffled to enhance the working of the model.Patient suffering from serious health care must be in frequent contact with health care experts to improve the confidence and learning about the improvement in their health.Generative medical AI give information to the patient in form of radiology imaging. Many conversation between patient's guardian and the medical expert are observed and the model is trained based on that.This will lead to clarity of the patient record and their confidence in medical issue.Each person were given a scenario to assume about their relative discharge summary and ask question .The manual evaluation were also done to see if the answer generated by the model satisfies the guardian.Novel answer from the model are evaluated to measure the factual correction and its relevance of answer by AI.

DATASET

Dataset have been created to perform based on synergy task,with effect of individuals to answer for the questions asked based on drug /treatment which is more than sum of their individual effect.The dataset is annotated in both english and spanish.It mainly focus on two major task of semantic indexing of QA and then next one is related to the answering of Question Answer.The questions of this dataset[19] has been taken from Articles of three batches,where B1 contains 34418 ,B2 contains 34711 and last batch B3 contains 31428 questions.

The snippets for QA is derived from two phases(i) Phase A(Retrieval of required Information),1000 questions are released and the participants are asked to answer this based on the reference from medical related fields.(ii)Phase B(Answering the question),accurate answers are expected from the participants based on the entity names.

From the Socrates period Community based question answering system has given a great impact on the social people to clear their doubts.The novice user may ask their queries in a public forum and it will be answered by the field experts or professional.Sometimes the repeated questions will be answered which is time consuming.This dateset[20] flags the repeated question as duplicate and tag it to the already answered forum so that it avoids manual annotation.In previous work the content was extracted from web .since the data are dynamic and tend to change it is hard to produce the exact answer for the queries addressed.The major limitation with this dataset is that when duplicate questionas are flagged it includes the unasnwered question also which may lead to loss of information for the user who posted the queries.

In EduQG,3,397 samples[21] are generated for the given sample content.This datasets is used for question distractor and question generation.The questions are annotated by educational experts and crowd workers.The big challenge is to develop a unique multiple choice question generator that is specially annotated by crowd workers.The sample content is converted to interrogative sentence and then feature engineering is applied to give a syntactic tree using the parser.The named Entity relation are extracted using the keyword which is helpful for question generation.All the Question generation follows two phases of generation.In phase (i),Natural and Macro Questions are undergoing information seeking technique to know from where the answers are documented.In phase (ii),the knowledge probing fall out to test and validate the knowledge of another person.

All the existing datasets fail to give answer directly in single hop i.e to extract answer from a single document or content. The model requires multiple document to answer as the question are diverse and not specific to any particular domain or it cannot be answered from pre existing knowledge. The flow of extracting answer is referred in Fig.1.

Fig.1.Phases of HotpotQA

To make the model efficient and time consuming ,it must be able to extract the answer form the available resource in limited time.Instead of retrieving the answer form the paragraph , a knowledge graph[22] is generated with tree structure for searching and retrieving the answer efficiently.Developing this datasets focus on generating answer to the natural question in real time about the health status,treatment recommendation.It build a knowledge graph[23] based on medical domain to keep the information organized with entities as root and attributes as leaves.The main focus is to concentrate on the question type and then utilize it for the comparison during processing of different set of question in knowledge tree.This system supports merging and re ranking candidate answers that are retrieved from different source.From the knowledge tree the medical entity has to be identified and the associated contents from the entity will be retrieved.Based on the ranking the best answer will be selected as the candidate answer.

The consumer based queries about health or medication is answered from the multiple source.The question are not synthetic ,all of them are taken from real forum where consumers queries are answered.In total 674 question are considered for annotation,the respective scores are also generated.The human annotated question's performance was evaluated by either RNN or CNN to identify question type.The initial stage of data creation is done by selecting consumer question about drugs and corresponding selecting the reference answer.In next step the baseline method of focusing the question type and recognition of answer is done.

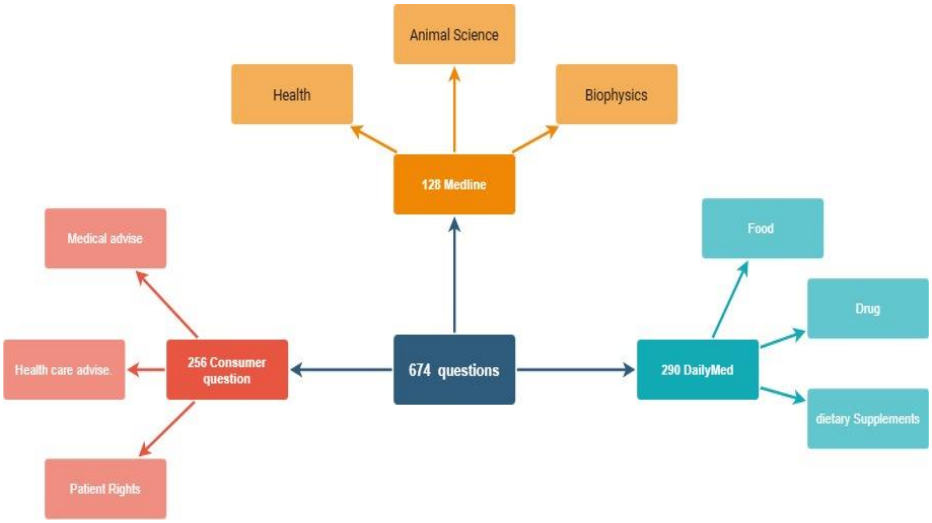


Fig.2.Question collected from different Sources

Dataset[24] that is used for answering complex question based on knowledge graph The biggest challenge of giving correct answer on domain question is solved in this datasets by matching the pairs in knowledge graph.The natural language questions are converted into triples by using a set of logical rules.This triples are arranged in order by WordNet and Marking Algorithm.The SimCSE algorithm[25] matches the triples in question along with triples in knowledge graph to retrieve the correct answer.Based on the type of content the knowledge as mentioned in Fig.2 is splited as generic and domain specific knowledge graph(it is insufficient to answer common sense knowledge.Apart from text some non textual information is also used for retrieving the answer as video information,image information are generally added to textual content.

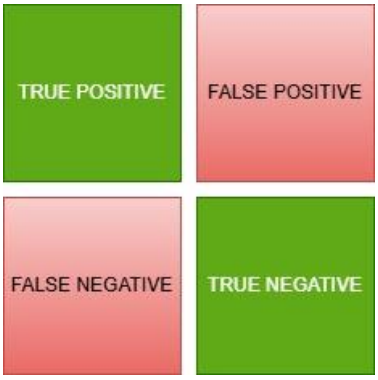
EVALUATION METRICS

There is no specific evaluation metrics for Question answering system, any metric that validate machine translated text is been used to evaluate the text. Manual observation is very tedious as it will take lot of time and effort to analysis the machine translated text.

Accuracy,is the most widely used evaluation metric for any machine learning and deep learning problems.It is calculated based on the number of answers answered correctly to the total number of correct answers.

ROGUE(Recall-Oriented Understudy for Gisting Evaluation) [26]has been used in most of the paper to evaluate the model.There are many types like ROUGE-N: This metric measures the overlap of n-grams (contiguous sequences of n words) between the generated summary and the reference summary. ROUGE-N typically considers unigrams (ROUGE-1), bigrams (ROUGE-2), and sometimes higher-order[27] n-grams (ROUGE-3, ROUGE-4, etc.)ROUGE-L: As discussed earlier, ROUGE-L measures the longest common subsequence between the generated summary and the reference summary. It focuses on surface-form similarity based on the longest sequence of words that appears in both texts.ROUGE-W: This metric is similar to ROUGE-L but incorporates weighted LCS (Longest Common Subsequence) measures, where matches at the beginning of the sequences are weighted more heavily.ROUGE-S: ROUGE-S computes the skip-bigram co-occurrence statistics between the generated and reference summaries. Skip-bigrams are pairs of words that occur in the same order in both texts but may have other words occurring between them.ROUGE-SU: ROUGE-SU extends ROUGE-S by considering skip-bigrams[28] with varying gap lengths. It's more flexible in capturing sentence-level structure and syntactic variation.ROUGE-WMD: This metric calculates the Word Mover's Distance between the generated and reference summaries. Word Mover's Distance is a measure of semantic similarity that considers the distances between word embeddings in a semantic space.

The harmonic mean of recall and accuracy yields the F1 score. A harmonic mean is a kind of average that is computed by multiplying the total number of values in a dataset by the reciprocal of each value in the datasets. The F1 score has a value between 0 and 1, where 1 represents a superior score.



1. Precision: Precision is the positive prediction accuracy. It computes the frequency with which the model accurately predicts positive values. It is calculated by dividing the total number of positive predictions (true positives plus false positives) by the number of genuine positive predictions.

$$\Pr ecision = \frac{Truepositive}{Truepositive + Falsepositive} \quad (1)$$

2. Recall (also known as True Positive Rate or Sensitivity): This measures how effectively a model can recognize real positive cases. It is calculated by dividing the total number of positive occurrences (true positives plus false negatives) by the number of genuine positive predictions. It assesses how well the model can account for every good example.

$$\text{Re}call = \frac{Truepositive}{Truepositive + Falsenegative} \quad (2)$$

This F1Score combines both Precision and Recall,

$$F1Score = \frac{2\Pr ecision * recall}{\Pr ecision + recall} \quad (3)$$

Model	Metric used	Result Achieved
Junda Wang(2024)	Accuracy	82.98%
Adnan Arefeen(2024)	ROGUE 1	Improved by 24.1%
Yanis Labrak(2023)	Exact Match Ratio	Improved by 18.6%
Yongping Du(2023)	F1 Score	79.80%
Debanjan Chaudhuri(2022)	F1 Score	83.60%
MahsaAbazari Kia(2022)	Exact Match Ratio	91.20%
Keqin Peng(2022)	Accuracy	93%
Pablo Schwarzenberg	F1 Score	90%

Table.2.Comparative analysis based on Result

As specified in Table.2 most Question Answering system are based on Large language model, BERTSCORE[29] is used as a relatively recent metric for evaluating the quality of machine-generated text, such as summaries or translations. It leverages pre-trained contextualized embeddings from models like BERT (Bidirectional Encoder Representations from Transformers) to compute similarity scores between generated and reference texts. BERTScore uses an F1 score as the primary evaluation metric. This score balances precision (the proportion of correctly identified relevant items among all items identified) and recall (the proportion of correctly identified relevant items among all relevant items). It provides a single numerical[30] value indicating the overall similarity between the generated and reference texts.BERTScore tokenizes sentences using the WordPiece tokenizer, which is also used in BERT models.

CONCLUSION

In conclusion, this survey paper provides a comprehensive overview of the methodologies, datasets, and evaluation metrics pertinent to Question Answering Systems (QAS), with a particular focus on Medical QAS. The paper delineates between open and closed domain QAS, as well as various types of QAS variants based on input and output, elucidating the diverse landscape within this field.Furthermore, the paper sheds light on the evaluation metrics commonly employed to assess the efficacy of QAS models, including precision, recall, accuracy, and F1 score, despite the absence of a specific metric tailored for QAS.By synthesizing insights from these sections, this survey paper provides valuable guidance for researchers and practitioners aiming to develop and refine QAS models. It not only underscores the importance of leveraging NLP techniques and curated datasets but also emphasizes the need for robust evaluation methodologies to gauge the performance of these systems accurately.

In summary, this survey paper serves as a road map for advancing research and development in the burgeoning field of Question Answering Systems, with a particular emphasis on the pivotal role of NLP in unlocking the potential of unstructured data across various domains, including medicine

AUTHOR CONTRIBUTIONS

Mrs. S. M. Keerthana: Conducted data collection, analysis, and material preparation. Drafted the initial manuscript.

Dr. K. Vijaya Kumar: Provided critical feedback on the manuscript, improving its quality and contributing to the final draft of the paper.

Dr Mohammad Nazmul Hasan Maziz: Reviewed the manuscript comprehensively, contributing to the theoretical framing and ensuring coherence across all sections. Assisted in finalizing the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFRENCES

[1] Basu, K., Varanasi, S. C., Shakerin, F., & Gupta, G. (2020). SQuARE: Semantics-based Question Answering and Reasoning Engine. Electronic Proceedings in Theoretical Computer Science, 325, 73–86. https://doi.org/10.4204/eptcs.325.13

[2] Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., & Xie, W. (2024). Towards Building Multilingual Language Model for Medicine. http://arxiv.org/abs/2402.13963

[3] Jiang, Z., Chi, C., & Zhan, Y. (2021). Research on Medical Question Answering System Based on Knowledge Graph. IEEE Access, 9, 21094–21101. https://doi.org/10.1109/ACCESS.2021.3055371

[4] Frisoni, G., Cocchieri, A., Presepi, A., Moro, G., & Meng, Z. (2024). To Generate or to Retrieve? On the Effectiveness of Artificial Contexts for Medical Open-Domain Question Answering. http://arxiv.org/abs/2403.01924

[5] Jiao, S., Zhu, Z., Wu, W., Zuo, Z., Qi, J., Wang, W., Zhang, G., & Liu, P. (2023). An improving reasoning network for complex question answering over temporal knowledge graphs. Applied Intelligence, 53(7), 8195–8208. https://doi.org/10.1007/s10489-022-03913-6

[6] Wang, Y., Lipka, N., Rossi, R. A., Siu, A., Zhang, R., & Derr, T. (2023). Knowledge Graph Prompting for Multi-Document Question Answering. http://arxiv.org/abs/2308.11730

[7] Kandula, V. V., & Bhattacharyya, P. (2023). Decision Knowledge Graphs: Construction of and Usage in Question Answering for Clinical Practice Guidelines. http://arxiv.org/abs/2308.02984

[8] Pramanik, S., Alabi, J., Saha Roy, R., & Weikum, G. (n.d.-a). UNIQORN: Unified Question Answering over RDF Knowledge Graphs and Natural Language Text. https://www.w3.org/RDF

[9] Aftabi, S. Z., Seyyedi, S. M., Maleki, M., & Farzi, S. (2024). ReQuEST: A Small-Scale Multi-Task Model for Community Question-Answering Systems. IEEE Access, 12, 17137–17151. https://doi.org/10.1109/ACCESS.2024.3358287

[10] Diomedi, D., & Hogan, A. (2021). Question Answering over Knowledge Graphs with Neural Machine Translation and Entity Linking. http://arxiv.org/abs/2107.02865

[11] Li, J., Wang, X., Wu, X., Zhang, Z., Xu, X., Fu, J., Tiwari, P., Wan, X., & Wang, B. (2023). Huatuo-26M, a Large-scale Chinese Medical QA Dataset. http://arxiv.org/abs/2305.01526

[12] Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Gasco, L., Krallinger, M., & Paliouras, G. (2021). Overview of BioASQ 2021: The ninth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. https://doi.org/10.1007/978-3-030-85251-1_18

[13] Hoogeveen, D., Verspoor, K. M., & Baldwin, T. (2015). CQADupStack: A benchmark data set for community question-answering research. ACM International Conference Proceeding Series, 08-09-Dec-2015. https://doi.org/10.1145/2838931.2838934

[14] Hadifar, A., Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2022). EduQG: A Multi-format Multiple Choice Dataset for the Educational Domain. http://arxiv.org/abs/2210.06104.

[15] Yang, Y., Yu, J., Hu, Y., Xu, X., & Nyberg, E. (2017). CMU LiveMedQA at TREC 2017 LiveQA: A Consumer Health Question Answering System. http://arxiv.org/abs/1711.05789

[16] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. http://arxiv.org/abs/1809.09600

[17] Ben Abacha, A., Mrabet, Y., Sharp, M., Goodwin, T. R., Shooshan, S. E., & Demner-Fushman, D. (2019). Bridging the gap between consumers' medication questions and trusted answers. Studies in Health Technology and Informatics, 264, 25–29. https://doi.org/10.3233/SHTI190176

[18] Zhao, Z., Jiang, Y., Liu, H., Wang, Y., & Wang, Y. (2023). LibriSQA: A Novel Dataset and Framework for Spoken Question Answering with Large Language Models. http://arxiv.org/abs/2308.10390.

[19] Arefeen, M. A., Debnath, B., & Chakradhar, S. (2023). LeanContext: Cost-Efficient Domain-Specific Question Answering Using LLMs. http://arxiv.org/abs/2309.00841

[20] Cai, P., Yao, Z., Liu, F., Wang, D., Reilly, M., Zhou, H., Li, L., Cao, Y., Kapoor, A., Bajracharya, A., Berlowitz, D., & Yu, H. (2023a). PaniniQA: Enhancing Patient Education Through Interactive Question Answering. http://arxiv.org/abs/2308.03253

[21] Cai, P., Yao, Z., Liu, F., Wang, D., Reilly, M., Zhou, H., Li, L., Cao, Y., Kapoor, A., Bajracharya, A., Berlowitz, D., & Yu, H. (2023b). PaniniQA: Enhancing Patient Education Through Interactive Question Answering. http://arxiv.org/abs/2308.03253

[22] Cao, X., Liu, Y., & Sun, F. (2023). Predict, pretrained, select and answer: Interpretable and scalable complex question answering over knowledge bases. Knowledge-Based Systems, 278. https://doi.org/10.1016/j.knosys.2023.110820

[23] Chen, H., Fang, Z., Singla, Y., & Dredze, M. (2024a). Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. http://arxiv.org/abs/2402.18060

[24] Chen, H., Fang, Z., Singla, Y., & Dredze, M. (2024b). Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. http://arxiv.org/abs/2402.18060

[25] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022). MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. http://arxiv.org/abs/2203.14371

[26] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B. A. y, … Natarajan, V. (2023). Towards Expert-Level Medical Question Answering with Large Language Models. http://arxiv.org/abs/2305.09617

[27] Wang, J., Yang, Z., Yao, Z., & Yu, H. (2024). JMLR: Joint Medical LLM and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability. http://arxiv.org/abs/2402.17887

[28] Wang, Y., Zhou, Y., Duan, C., Bao, J., & Zhao, T. (2023). HopPG: Self-Iterative Program Generation for Multi-Hop Question Answering over Heterogeneous Knowledge. http://arxiv.org/abs/2308.11257

[29] Zhang, J., Zhang, H., Zhang, D., Liu, Y., & Huang, S. (2023). End-to-End Beam Retrieval for Multi-Hop Question Answering. http://arxiv.org/abs/2308.08973

[30] Zhang, M., Ma, Y., Li, Y., Zhang, R., Zou, L., & Zhou, M. (2023). Two is Better Than One: Answering Complex Questions by Multiple Knowledge Sources with Generalized Links. http://arxiv.org/abs/2309.05201.