IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

A Review on Conversational Question Answering (CQA)

Mr. Jeevan Tonde[1] and Dr. Satish Sankaye[2]
Dr. G. Y. Pathrikar College of Computer Science & Information Technology, Aurangabad, Maharashtra, India
MGM University, Aurangabad, Maharashtra, India

Abstract: *The internet is constantly changing how people communicate with one another, manage their daily lives, and share the information. As a result of digitization, automated Question Answering Systems are becoming more important in extracting useful information from the knowledge sources. In general, Question Answering (QA)mechanisms are intended to answer a specific question only once (so-called single-turn).However, to satisfy a user's information needs, the Conversational Question Answering (CQA) system must grasp the given context and engage in multi-turn QA. While single-turn QA leads the majority of current research, multi-turn QA has rapidly gained prominence due to the availability of large-scale multi-turn QA datasets and the development of "pre-trained language models". In this research paper, we have presented a comprehensive review on the Conversational Question Answer (CQA) field which includes need of CQA followed by categorization of QA system, list of available datasets and various evaluation metrics.*

Keywords: Question Answering, Conversational Question Answering, Natural language processing

I. INTRODUCTION

Due to developments in learning technologies, question-answering has recently attracted a great deal of interest from the artificial intelligence community. Question answering is the process of automatically responding to human-posed natural language queries. Conversational Question Answering (CQA)[1] is where the given context must be understood by a system. i.e., designing Conversational AI systems that not only match or exceed a human's ability to carry on an engaging conversation, but also deliver solutions to a wide range of questions. One of the prominent objectives in the field of artificial intelligence has been to produce information that ranges from recent news about sports to a biography of a well-known political figure.

Recently, conversational interfaces/assistants such as Amazon's Alexa, Apple's Siri, Google's Google-Assistant, and others[2] have become a focal point in both academic and industry research because of their rapid market uptake and rapidly increasing range of capabilities.The previous generation of these assistants concentrated on short, task-oriented discussion, such as playing music or asking for theparticular information.They have not focused on longer free-form dialogues that naturally occur in general human interaction.

The conversational AI domain can be divided into three categories:

(I) "Task-oriented dialogue systems" are required to complete the tasks on behalf of users. e.g., "booking a flight ticket" or "playing a song from movie"

(II) "Chat-oriented dialogue systems" that conducts a natural and interactive dialogue with users. e.g., "Do you like apple?" or "What is your name?"

(III) "Question Answering (QA) dialogue systems thatgives clear and concise replies to user's questions based on data derived from various data sources such as "text documents or knowledge stores".

1.1 Difference between CQA and QA

CQA's function is different from traditional QAs in a few aspects. Traditional Question Answering (QA) systems have questions that are independent to one another and are based on the passage that is presented. Contrarily, CQA questions are linked to one another, which presents a completely distinct set of challenges.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

In CQA the model must encode nor only thecurrent question and source paragraph but also the prior history turns in order to find the right response for the question. More particularly, as demonstrated in figure 1, Questions 2 and Question 3 are connected to Question 1.

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Fig. 1.: A Sample from CoQA dataset[3]

II. CLASSIFICATION OF CQA SYSTEM

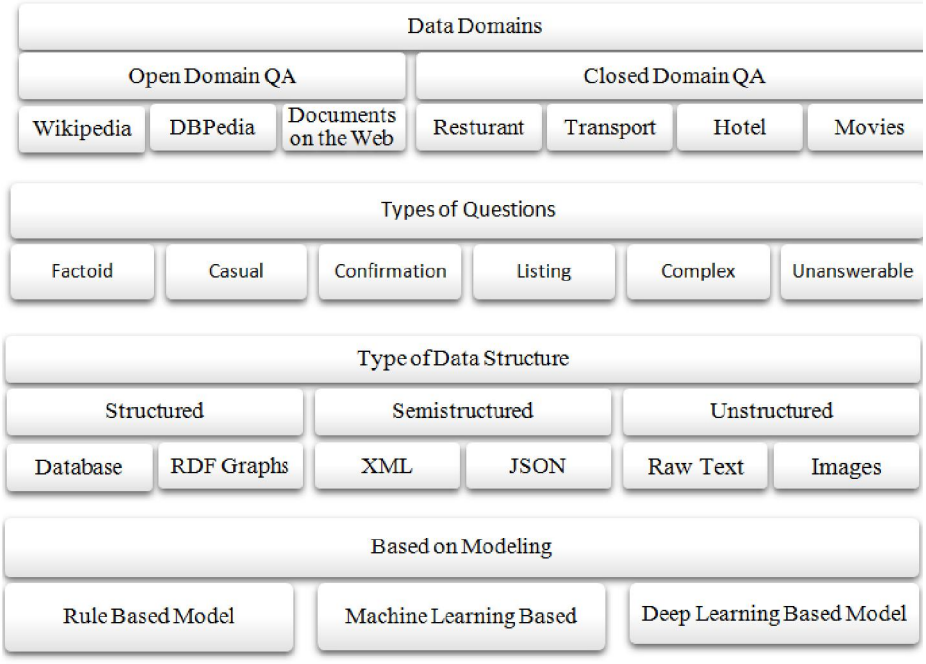

Fig. 2: Classification Conversational Question Answering System[4]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

2.1 Based on data domains

2.2.1 Open domain Question answering

Open domain questions[5]are domain-free such as Wikipedia anddocuments available on the Web. The question repository for Open domain is large as compared to closed domain question answering.

2.2.2 Closed domain Question Answering

Closed or Restricted domain Question answering is restricted to specific application domains such as travel, restaurants, movie and Clinics. Closed domain question answering has a smaller question repository than open domain question answering.

2.2 Based on the types of question

In this section we discussedthe details of the classification of Question Answering (QA) systems along witha description of the classification. The job of producing answers to questions is associated with the type of questions asked.

2.2.1 Factoid Question

The factoid type questions are generally begun with a wh-word,e.g., what, where, which, when, how, etc. These are straightforward, fact-based questions that require solutions in a single sentence or brief phrase. Factoid type questions can be answered by short phrases, such as persons, dates and locations

2.2.2 Casual Question

Causal questions are asked by users who want replies as justifications, explanations, elaborations, and so on about certain objects or events. Causal questions require descriptive responses that can range from sentences to paragraphs to a whole document.

2.2.3 Confirmation Question

Confirmation questions require yes or no replies. Answering confirmation questions requires world knowledge, an inference system, and common-sense reasoning.

2.2.4 Listing Question

List-type questions, such as "Names of movies released in 2017," call for a list of facts or things to be mentioned as the solution. The response types for list type inquiries are named entities.

2.2.5 Complex Question

Complex questions are more difficult to answer and often require inferring and synthesizing information from multiple documents to get multiple nuggets as answers.

2.3 Based on type of data source

Conversational Question Answering (CQA)systems are classified according to the data sources they use to give an answer.

2.3.1 Structured Data Source

In a structured document, data is preserved as entities. A table entity can be associated with several properties. The definition of these attributes is called as metadata, and it is stored in a schema.A query language is used to gain access to the data and extract crucial information from the schema

Examples of structured datasets are Resource Description Framework (RDF) graphs, Question Answering over Linked Data (QALD)[6]Footnote and LC-QuAD

2.3.2 Semi-structured Data Source

Semi-structured datais information that doesn't consist of structured data (relational database) but still has some structure to it. The data-format can be HTML, XML, JSON, NSQL.

Examples of structure datasets are TabMCQ[7] and Question Answering using Semi-structured Metadata (QuaSM)[8]

2.3.3 Unstructured Data Source

The data which is stored in the unstructured data sources can be of any type and requires the use of information retrieval or natural language processing techniques to find out the relevant answer. When compared to structured data sources, the reliability of getting the correct answers is poor. Examples of unstructured datasets are SQuAD, Question Answering in Context (QuAC)[9] , and CNN & Daily Mail.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

2.4 Classification based on Modelling approach

2.4.1 Rule based Model

These systems require rule sets that define paths for questions based on the question type. The path taken by the answer extraction process for "where is the Gateway of India?" would be different from "who is the Prime Minister of India". Syntactic analysis, morphological analysis, Part-Of-Speech tagging, and Named Entity Recognition were eventually added to these systems to improve response matching.

2.4.2 Machine-learning based Model

In this model for categorization, the question's primarily processed result is fed into machine learning models such as support vector machine (SVM), decision tree (DT), and nave-bayes (NB).

2.4.3 Deep Learning based Model

The development of QA research is moving away from just "machine learning-based models" and toward deep learning-based models as a result of the accessibility of computing power and the emergence of "recurrent neural network (RNN)" based models in the text processing sector. For a given question answering task, finetuning of pretrained transformer models like Google's BERT [10] is the current state of the art.

III. QUESTION ANSWERING POPULAR DATASETS

3.1 SQuAD

The Stanford Question Answering Dataset (SQuAD)[11] is a dataset for reading comprehensionwhich includes questions posed by people on a set of Wikipedia articles, and the answer to every question is a segment of text, from the corresponding reading passage, or the question might be not able to answer. "SQuAD 2.0" data contains more than 100,000 questions.

3.2 Natural Questions (NQ)

The Natural Questions[12] corpus is a question answering dataset containing 300,000 training examples, 7,842 test examples and 7,830 development examples. Each example includes a Google.com query and a related Wiki page. Each Wiki page contains one or more short spans from the annotated text that include the actual solution, as well as a paragraph (or lengthier answer) on the page that answers the query.

3.3 Question Answering in Context (QuAC)

Question Answering in Context (QuAC) [9]is a dataset for modeling, understanding, and participating in information seeking dialog. In this dataset, instances are composed of an interactive dialogue between two crowd workers: a student who asks a series of freeform questions in order to learn as much as possible about a hidden Wikipedia text, and a teacher who responds to the questions by providing short excerpts (spans) from the text. It comprises 14K information-seeking QA dialogues and 100K QA pairs in total.

3.4 HOTPOTQA

HOTPOTQA[13] is a dataset which contains 113k Wikipedia-based question-answer pairs with four key features. The questions are diverse and not constrained to any pre-existing knowledge bases or knowledge schemas, sentence-level supporting facts required for reasoning, allowing QA systems to reason with strong supervision and explain predictions, and a new type of factoid comparison questions to test QA systems' ability to extract relevant facts and perform necessary comparison.

3.5 ELI5

ELI5 (Explain Like I'm Five)[14] is a longform question answering dataset. It is a large-scale, high-quality data set that includes web articles and two pre-trained algorithms. Facebook generated the dataset, which includes 270K conversations with diverse, open-ended inquiries that demand multi-sentence answers.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

3.6 ShARC

Shaping Answers with Rules through Conversations (ShARC)[15] is a QA dataset which requires logical reasoning, elements of entailment/NLI and natural language generation. The dataset includes 32k task instances based on real-world regulations as well as crowd-sourced queries and scenarios.

3.7 MS MARCO

MS MARCO[16] or Human Generated MAchine Reading COmprehension Dataset is a large-scale dataset created by Microsoft AI & Research. The dataset includes 1,010,916 anonymized questions from Bing's search query records, each with a human-made answer and 182,669 totally human rewrote produced answers. This dataset is mainly intended for non-commercial research purposes only to promote advancement in the field of artificial intelligence and related areas

3.9 TWEETQA

TWEETQA[17]is a social media-focused question answering dataset. This dataset was generated by IBM and University of California researchers and is the first large-scale dataset for QA over social media data. There are now 10,898 articles, 17,794 tweets, and 13,757 crowdsourced question-answer pairings in the dataset.

3.10 NEWSQA

NewsQA[18] is a challenging machine comprehension dataset of over 100,000 human-generated question-answer pairs. The dataset was compiled by crowd-workers who supplied questions and answers based on a set of over 10,000 CNN news articles, with answers consisting of text spans from the respective articles. The dataset includes 119,633 natural language questions asked by crowdworkers on 12,744 CNN news stories.

IV. EVALUATION METRICS FOR QUESTION ANSWERING

There are several different evaluation metrics available in the field of question-answering[19][4]
Some of evaluation metrics are

1. Precision
2. Recall
3. F1 score
4. Rouge
5.BLEU

4.1 Precision, Recall, F1 Score

For the Multiple choice-based Question Answering, standard way to measure performance is to calculate precision, recall or F1 score [21] as given by next equations.

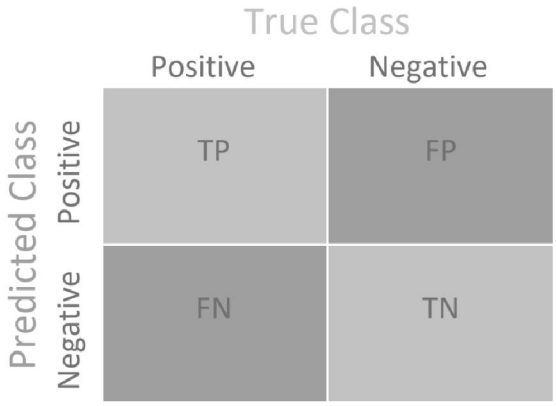

Fig. 3: Confusion Matrix

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

$$Precision = \frac{Total\ Positive}{Predicted\ Positive}$$

$$Recall = \frac{True\ Positive}{Actual\ Positive}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.2 Rouge

ROUGE ("RecallOriented Understudy for Gisting Evaluation") is a set of metrics for assessing automatic summary of literature and machine translations.It compares a self-generated summary or translation against a set of reference summaries(typically human-produced). Assume you have the system and reference summaries listed below.

$$Recall = \frac{Number\ of\ overlapping\ words}{Total\ words\ in\ reference\ summary}$$

$$Precision = \frac{Number\ of\ overlapping\ words}{Total\ overlapping\ words\ in\ system\ summary}$$

ROUGE-N measures unigram, bigram, trigram and higher order n-gram overlap

4.3 BLEU Data set

BLEU (Bilingual Evaluation Understudy)[20]It is a method for determining the precision of machine-translated text from one natural language to another. The theory behind Quality, according to BLEU, is defined as the resemblance between a machine's output and that of a human:"the closer a machine translation is to a professional human translation, the better it is". BLEU was one of the first measurements to demonstrate a strong relationship with human judgements of quality, which is now one of the most widely used automated and affordable metrics.

Individual translated segments—typically sentences—have their scores determined by comparison to a collection of accurate reference translations. To determine an approximation of the overall quality of the translation, these scores are then averaged throughout the entire corpus. Grammar accuracy and intelligibility are not taken into consideration.

V. OBSERVATIONS

Conversational Question Answering (CQA) systems have emerged as a crucial technology for decreasing the interactional gap between machines and humans as a result of advances in pre-trained language modelling and the introduction of conversational datasets. This advancement simplifies and accelerates the domains such as online customer service, interactions with IoT devices in smart spaces, and search engines, allowing CQA to achieve its social and economic implications. The primary issues are the successful absorption of contextual information, the ability to infer inquiries, and the ability to offer efficient clarifying questions.The educational sector is one of the areas that can greatly profit from the use of Conversational AI. It can increase productivity, communication and learning, provide effective teaching support, and reduce ambiguity.

VI. CONCLUSION

In this article, we have provided a survey of the current cutting-edge Conversational Question Answering (CQA) System, based on the type of questions, different research directions. We have also studied various datasets available for conversational Question answeringand various evaluation measures. We believe that this review article will act as the researchers' essential guide and open the door to future efforts to streamline the study of conversational question-answering.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

REFERENCES

[1] A. Bansal, Z. Eberhart, L. Wu, and C. McMillan, "A Neural Question Answering System for Basic Questions about Subroutines." arXiv, Jan. 11, 2021. Accessed: Oct. 09, 2022. [Online]. Available: http://arxiv.org/abs/2101.03999

[2] G. Terzopoulos and M. Satratzemi, "Voice Assistants and Smart Speakers in Everyday Life and in Education," *Inform. Educ.*, pp. 473–490, Sep. 2020, doi: 10.15388/infedu.2020.21.

[3] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A Conversational Question Answering Challenge." arXiv, Mar. 29, 2019. doi: 10.48550/arXiv.1808.07042.

[4] H. A. Pandya and B. S. Bhatt, "Question Answering Survey: Directions, Challenges, Datasets, Evaluation Matrices." arXiv, Dec. 07, 2021. doi: 10.48550/arXiv.2112.03572.

[5] P.-M. Ryu, M.-G. Jang, and H.-K. Kim, "Open domain question answering using Wikipedia-based knowledge model," *Inf. Process. Manag. Int. J.*, vol. 50, no. 5, pp. 683–692, Sep. 2014, doi: 10.1016/j.ipm.2014.04.007.

[6] A. Perevalov, D. Diefenbach, R. Usbeck, and A. Both, "QALD-9-plus: A Multilingual Dataset for Question Answering over DBpedia and Wikidata Translated by Native Speakers." arXiv, Feb. 07, 2022. doi: 10.48550/arXiv.2202.00120.

[7] "(PDF) TabMCQ: A Dataset of General Knowledge Tables and Multiple-choice Questions." https://www.researchgate.net/publication/301846039_TabMCQ_A_Dataset_of_General_Knowledge_Tables_and_Multiple-choice_Questions (accessed Oct. 09, 2022).

[8] D. Pinto *et al.*, "QuASM: a system for question answering using semi-structured data," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, New York, NY, USA, Jul. 2002, pp. 46–55. doi: 10.1145/544220.544228.

[9] E. Choi *et al.*, "QuAC : Question Answering in Context." arXiv, Aug. 27, 2018. doi: 10.48550/arXiv.1808.07036.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text." arXiv, Oct. 10, 2016. doi: 10.48550/arXiv.1606.05250.

[12] T. Kwiatkowski *et al.*, "Natural Questions: A Benchmark for Question Answering Research," *Trans. Assoc. Comput. Linguist.*, vol. 7, pp. 453–466, Nov. 2019, doi: 10.1162/tacl_a_00276.

[13] Z. Yang *et al.*, "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct. 2018, pp. 2369–2380. doi: 10.18653/v1/D18-1259.

[14] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, "ELI5: Long Form Question Answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 3558–3567. doi: 10.18653/v1/P19-1346.

[15] Y. Gao, J. Li, C.-S. Wu, M. R. Lyu, and I. King, "Open-Retrieval Conversational Machine Reading." arXiv, Nov. 24, 2021. Accessed: Oct. 09, 2022. [Online]. Available: http://arxiv.org/abs/2102.08633

[16] P. Bajaj *et al.*, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset." arXiv, Oct. 31, 2018. doi: 10.48550/arXiv.1611.09268.

[17] W. Xiong *et al.*, "TWEETQA: A Social Media Focused Question Answering Dataset." arXiv, Jul. 14, 2019. doi: 10.48550/arXiv.1907.06292.

[18] A. Trischler *et al.*, "NewsQA: A Machine Comprehension Dataset." arXiv, Feb. 07, 2017. doi: 10.48550/arXiv.1611.09830.

[19] M. Barbella, M. Risi, and G. Tortora, "A Comparison of Methods for the Evaluation of Text Summarization Techniques:," in *Proceedings of the 10th International Conference on Data Science, Technology and Applications*, Online Streaming, --- Select a Country ---, 2021, pp. 200–207. doi: 10.5220/0010523002000207.

[20] "BLEU - a Hugging Face Space by evaluate-metric." https://huggingface.co/spaces/evaluate-metric/bleu (accessed Oct. 09, 2022).