Knowledge and Information Systems (2022) 64:3151–3195
https://doi.org/10.1007/s10115-022-01744-y

SURVEY PAPER

Conversational question answering: a survey

Munazza Zaib[1] · **Wei Emma Zhang**[2] · **Quan Z. Sheng**[1] · **Adnan Mahmood**[1] · **Yang Zhang**[1]

Received: 28 May 2021 / Revised: 29 July 2022 / Accepted: 31 July 2022 /
Published online: 6 September 2022

Abstract
Question answering (QA) systems provide a way of querying the information available in various formats including, but not limited to, unstructured and structured data in natural languages. It constitutes a considerable part of conversational artificial intelligence (AI) which has led to the introduction of a special research topic on *conversational question answering* (CQA), wherein a system is required to understand the given context and then engages in multi-turn QA to satisfy a user's information needs. While the focus of most of the existing research work is subjected to single-turn QA, the field of multi-turn QA has recently grasped attention and prominence owing to the availability of large-scale, multi-turn QA datasets and the development of pre-trained language models. With a good amount of models and research papers adding to the literature every year recently, there is a dire need of arranging and presenting the related work in a unified manner to streamline future research. This survey is an effort to present a comprehensive review of the state-of-the-art research trends of CQA primarily based on reviewed papers over the recent years. Our findings show that there has been a trend shift from single-turn to multi-turn QA which empowers the field of Conversational AI from different perspectives. This survey is intended to provide an epitome for the research community with the hope of laying a strong foundation for the field of CQA.

Keywords Question answering · Conversational agents · Conversational machine reading comprehension · Knowledge base · Conversational AI

✉ Munazza Zaib
munazza-zaib@hdr.mq.edu.au

✉ Quan Z. Sheng
michael.sheng@mq.edu.au

[1] School of Computing, Faculty of Science and Engineering, Macquarie University, Sydney, NSW 2109, Australia

[2] School of Computer Science, The University of Adelaide, North Terrace, Adelaide, SA 5005, Australia

1 Introduction

Designing an intelligent dialog system that not only matches or surpasses a human's level on carrying out an interactive conversation, but also answers the questions on a variety of topics, i.e., ranging from recent news about National Aeronautics and Space Administration (NASA) to a biography of a famous political leader, has been one of the outstanding goals in the field of artificial intelligence (AI) [22]. A quickly increasing number of research papers prove the promising potential and the growing interest of researchers from both academia and industry in conversational AI.

Conversational AI constitutes an integral part of natural user interfaces [22] and is attracting significant attention from researchers in information retrieval (IR), natural language processing (NLP), and deep learning (DL) communities. For example, AAAI 2020 introduced a special workshop focusing on "Reasoning for Complex Question Answering," which featured a special focus on machine intelligence and common sense reasoning. Similarly, SIGIR 2018 introduced a new track entitled "Artificial Intelligence, Semantics and Dialog" to bridge the gap between IR and AI. The track is especially focused on QA, conversational dialog agents, and deep learning for IR and agents. One of the top conferences in NLP, EMNLP, has had a track called "Information Retrieval and Question Answering" for years, and from 2019, it has started inviting papers for the field of "Question Answering" as a separate track owing to the increasing research interests of the community and its faced-paced growth.

The field of conversational AI can be segregated into three groups, namely (i) *task-oriented dialog systems* that are required to perform tasks on the users' behalf such as making a reservation in a restaurant or scheduling an event, (ii) *chat-oriented dialog systems* that need to carry out a natural and interactive conversation with the users, and (iii) *QA dialog systems* that are responsible to provide clear and concise answers to the users' questions based on information deduced from different data sources such as text documents or knowledge bases. The examples of each of the aforementioned categories are given in Fig. 1. The conversation shown in Fig. 1 comprises of multiple turns and each turn consists of a question and an answer [77].

The chat-oriented and task-oriented dialog systems have been well-researched topics resulting in a number of successful dialog agents such as Amazon Alexa[1], Apple Siri[2], and Microsoft Cortana[3]. However, QA dialog systems are fairly new and still require extensive research. Many QA challenges have been identified and initial solutions have been proposed [2, 17, 34, 36, 75, 94, 98, 114], giving the rise of *Conversational Question Answering* (CQA). CQA techniques form the building blocks of QA dialog systems. The idea behind CQA is to ask the machine to answer a question based on the provided passage and this, in turn, has the potential to revolutionize the way humans interact with the machines. However, this interaction could turn into a multi-turn conversation if a user requires more detailed information about the question. The notion of CQA can be thought of as a simplified but concrete conversational search setting [68], wherein the system returns one correct answer to a user's question instead of a list of relevant documents or links as is the case with traditional search engines. The top search engine companies such as Microsoft and Google have incorporated CQA into their mobile-based search engines (also known as *digital assistants*) to improve the users' experience when interacting with them.

[1] https://www.amazon.com.au/b?node=5425666051.

[2] https://www.apple.com/au/siri/.

[3] https://www.microsoft.com/en-us/cortana.

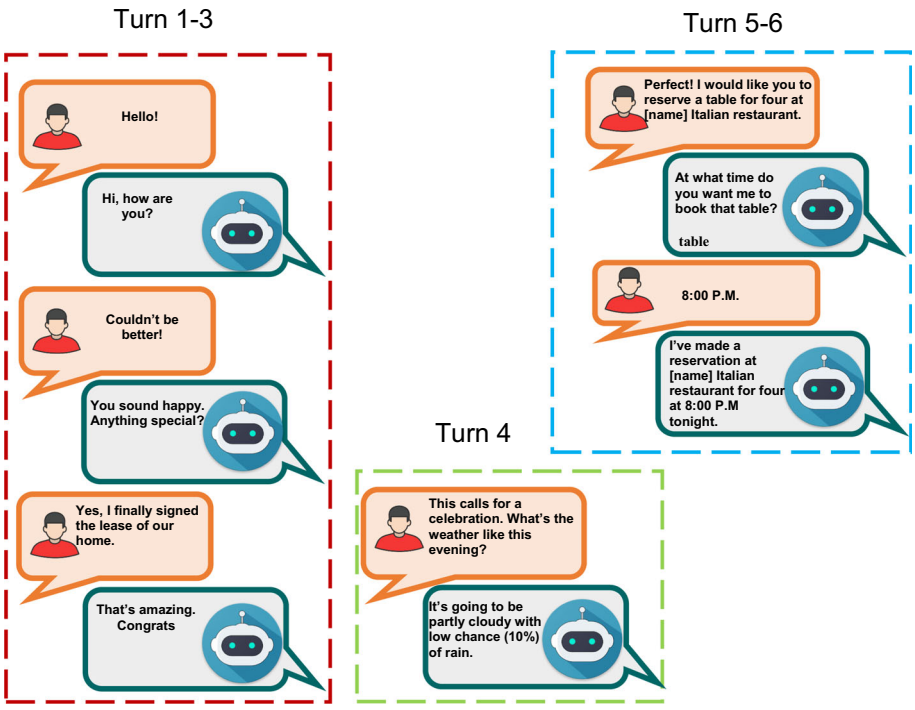


Fig. 1 Categorizations of conversational AI. Turn 1–3 depict chat-oriented dialog system, turn 4 portrays the element of QA dialog system, and turn 5–7 reflect depicts the task-oriented conversation

CQA is an effective way for humans to gather information and is considered as a benchmark task to evaluate a machine's capability to understand and comprehend the input provided in written natural language [117]. Such CQA systems have significant applications in areas like customer service support [16] or QA dialog systems [25, 77]. The task of CQA poses several challenges to the researchers hence resulting in considerable interesting yet innovative researches over the past few years.

1.1 Papers' selection

The research papers reviewed in this survey are high-quality papers selected from the top NLP and AI conferences, including but not limited to, ACL,footnotehttps://www.aclweb.org/. SIGIR,[4] NeurIPS,[5] NAACL,[6] EMNLP,[7] ICLR,[8] AAAI,[9] IJCAI,[10] CIKM,[11] SIGKDD,[12] and

[4] https://sigir.org/.

[5] https://nips.cc/.

[6] https://naacl.org/.

[7] https://sigdat.org/.

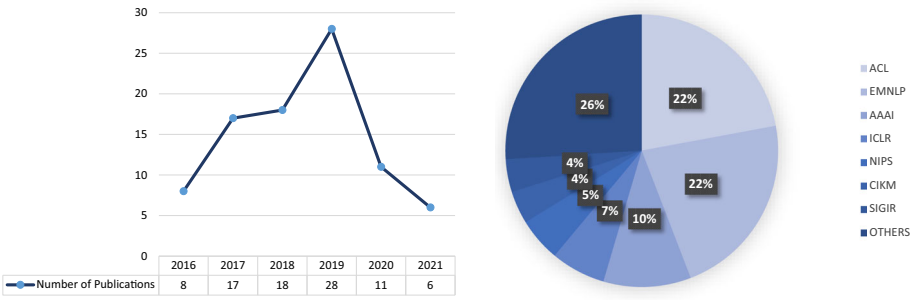
[8] https://iclr.cc/.

[9] https://www.aaai.org/.

[10] https://www.ijcai.org/.

[11] http://www.cikmconference.org/.

[12] https://www.kdd.org/.



(a) Year-wise distribution. **(b)** Venue distribution.

Fig. 2 **a** Year-wise statistics of the selected survey papers between 2016 and 2021 inclusive. The figure depicts that the field of CQA saw its rise recently. **b** Venue-wise distribution of the reviewed research works

WSDM.[13] Other than published research papers in the aforementioned conferences, we have also considered good papers in e-Print archive[14] as they manifest the latest research outputs. We selected papers from archive using three metrics: paper quality, method novelty, and the number of citations (optional).

Figure 2a depicts the year-wise distribution of the papers reviewed in our survey. Our survey encompasses over 80 top-notch conferences and journal papers. The number of papers pertinent to CQA steadily increases from year 2016 onwards, with the highest being in 2019. Coincidentally, 2019 also marks the year when the fields of natural language generation and natural language understanding were revolutionized with the introduction of pre-trained language models. These pre-trained language models have the potential to address the issue of data scarcity and bring considerable advantages by generating contextualized word embeddings [112]. This rise of interest depicts the gradual shift in focus of the researchers in both academia and industry in utilizing pre-trained language models for the design of CQA systems. Also, Fig. 2b portrays the venue-wise distribution of the research works we have reviewed, with ACL and EMNLP being the top venues for natural language-related progress. We note, though, that more than 25% of papers come from a variety of conferences/journals outside of the typical venues further attesting to the fact that this is an interdisciplinary topic spanning different areas such as knowledge management, knowledge discovery, information retrieval, and artificial intelligence.

1.2 What makes this survey different?

There have been several published literature reviews on QA systems, i.e., in the context of both machine reading comprehension (MRC) and knowledge-based question answering (KB-QA). In [22], the authors provide an overview of Conversational AI with a detailed discussion of neural methods and deep learning techniques being used in designing efficient conversational agents. These conversational agents include task-oriented dialog systems, chat-oriented dialog systems, and QA dialog systems. Although the paper sheds some light on several research works and datasets pertinent to CQA, it does not cover the recent trends and methods on CQA. The authors in [21] recently published their literature review which

[13] http://www.wsdm-conference.org/.

[14] https://arxiv.org/.

primarily highlights the complex QA over knowledge bases. The paper covers all the datasets and different approaches that are employed in complex KB-QA systems along with the discussion of complex QA. CQA is just mentioned as a "future trend" with minimal discussion. A summary of the techniques and methods of single-turn QA is presented in [42] along with proposing a general modular architecture needed for it. The paper further discusses the techniques that could be used in each module. Again, CQA is discussed briefly as a newly emerging trend along with the different challenges. The recent trends in pre-trained language modeling and their applications in dialog systems are discussed briefly in [112]. The short survey focuses on utilization of these models specially in QA systems and generally in task-oriented and chat-oriented dialog systems. However, the survey lacks the discussion on architectures based on traditional or flow-based models for CQA. Another recent effort [24] delineates on the latest trends and methods to cater to the successful implementation of multi-turn MRC. However, it lacks the discussion on other forms of multi-turn QA.

The key aspect that makes this survey to stand out among its predecessors is its focus on CQA encompassing both sequential KB-QA and conversational machine reading comprehension (CMRC). Multi-turn QA has been discussed very nominally in previous surveys. It is an essential aspect to consider when discussing the process of carrying out a natural conversation with a machine. Based on the review, we thoroughly discuss the research works of CQA, the techniques employed, and highlight the merits and demerits of different techniques. Finally, we highlight and discuss existing challenges related to the field of CQA along with an attempt to suggest some application areas.

The field of CQA is witnessing its golden era in terms of research publications and this calls for having a strong background work that discusses its challenges and trends as a separate field than single-turn QA. Thus, this survey is an effort to establish a strong foundation for CQA which would benefit the research communities as well. This work provides detailed insights into important ideas pertinent to CQA systems that are needed to design interactive and engaging conversational systems. To the best of our knowledge, this is the first work to investigate the field of CQA in detail. We hope that this paper would turn out to be a valuable resource for researchers who are interested in this area.

1.3 The survey structure

The rest of the paper is organized as follows. Section 2 delineates on a brief background of single-turn QA and leads the discussion on CQA. This section further highlights the categorization of CQA systems based on the source they utilize to answer the questions. Section 3 describes the task of sequential KB-QA system and the general architecture it employs. The section further highlights the techniques used in each module of the system to effectively carry out the task of sequential QA. Section 4 describes the task of CMRC and how it differs from typical machine reading comprehension (MRC). The section also describes how the general architecture of MRC can be adapted for CMRC. It further describes the decomposition of the architecture in different modules and techniques employed in each of them. Section 5 describes the datasets introduced to further improve the work in the field of CQA along with a qualitative comparison of each of them. Section 6 highlights the potential applications of the CQA systems in commercial areas along with the research trends that should be explored to leverage the strength of these systems more effectively. Finally, Sect. 7 offers some concluding remarks.

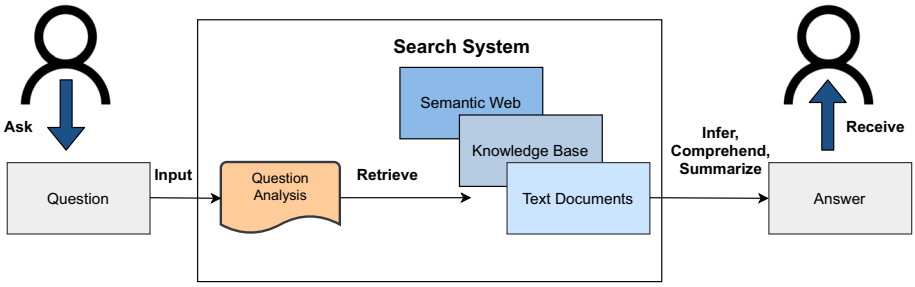


Fig. 3 High-level or generic architecture of QA systems where search system corresponds to the different sources. The specific architecture of a QA system depends on the underlying data source

2 Conversational question answering

Question answering in general involves accessing different data sources to find the correct answer for an asked question, as depicted in Fig. 3. It dates back to the 1960s [54] when early QA systems, due to rule-based methods and absurdly small size of available datasets, did not achieve well, thereby making it difficult to be used in practical applications. These systems saw their rise in 2015 and this largely was associated with two driving factors:

(a) The use of deep learning methods to capture the critical information in QA tasks that outperform the traditional rule-based models, and
(b) The availability of several large-scale datasets, i.e., Stanford Question Answering Dataset (SQuAD) [75], Freebase [5], **M**icrosoft **MA**chine Reading **CO**mprehension (MS MARCO) [57], DBpedia [39], and CNN & DAILY MAIL [56], which make it possible to deal with the task of QA on neural architectures more efficiently and further provide a test bed for evaluating the performance of these models.

To realize the QA tasks more close to the real-world scenarios, several advanced research directions have emerged recently. One such direction is CQA [42], which introduces a new dimension of dialog systems that combines the elements of both chitchat and QA. CQA is a *system ask, user respond* kind of setting where the system can asks a user multiple questions to understand the user's information need [115]. Usually, a user starts the conversation with a particular question in mind and the system searches its database to find an appropriate solution to that query. This could turn into a multi-turn conversation if the user needs to have more detailed information about the topic.

2.1 Categorization of CQA systems

There are several ways of structuring the different aspects of a QA system. Since CQA is categorized as a subcategory of QA, the same categorization can be used for CQA systems as well. The categorization of the CQA model could be realized on the basis of the data domain, types of questions, types of data sources, and the types of systems that we are building for the questions at hand [52]. Figure 4 manifests the possible options that could be utilized to structure a CQA system. The details of each of the category are given in the rest of this section.

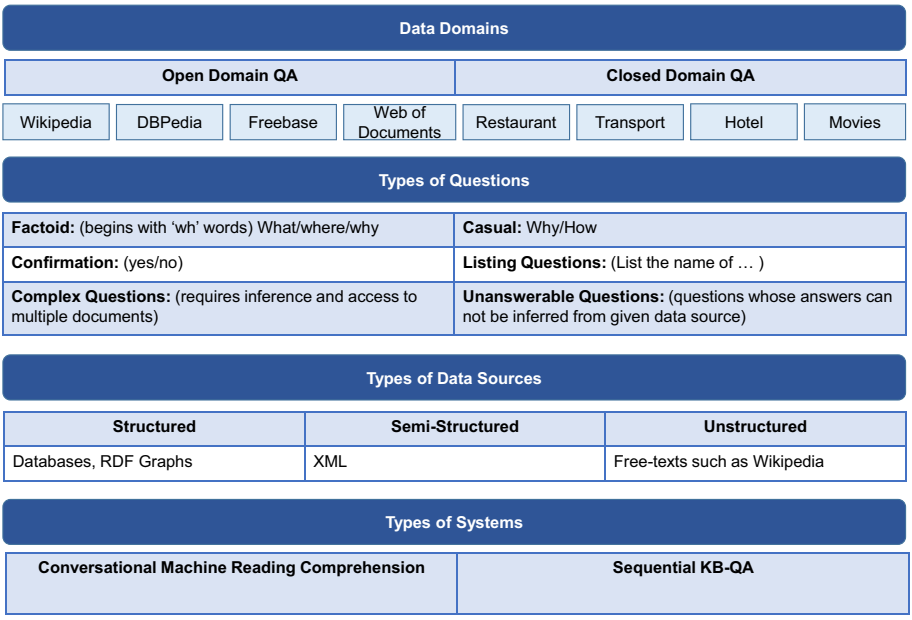


Fig. 4 Categorization of CQA on the basis of: (i) data domains, (ii) types of questions, (iii) types of data sources, and (iv) types of systems [52]

2.1.1 Data domains

Questions asked by users are either open-domain [33, 108] in which questions are domain-free and in a broad range, or restricted to specific application domains (i.e., closed-domain) such as Travel [3], Restaurants [8], Movies [9], and Hospitals [8]. The question repository of closed-domain question answering is smaller compared to open-domain question answering. This makes the models designed for closed-domain QA less transferable than the models for open-domain QA.

It should be noted that the subcategories of open-domain QA and closed-domain QA are the examples of generic and task-specific datasets.

2.1.2 Types of questions

Questions can be easily classified into various categories primarily depending upon their complexity, the nature of the response, or the techniques that should be utilized to answer them [52]. The classification based on the questions commonly asked by the users is delineated as follows:

Factoid questions Questions which expect the system to find a simple and fact-based answer in a short sentence, e.g., "*who acted as Chandler in FRIENDS?*." Factoid questions typically begin with a *wh*-word. Different extraction techniques can be employed to find the answers to the factoid questions. The techniques first recover latent or hidden information in the given question and then look for the answer in the given text using either structure matching [89] or reasoning [31]. FreebaseQA [33] is one of the examples of factoid QA dataset.

Confirmation questions Questions which require the answer in a binary format, i.e., yes or no, e.g., "*Is Sydney the capital of Australia?*." As the answers are not simple extractive text spans from the given source, a strong inference mechanism is needed to deduce the answers of confirmation type questions [52]. While there may be a lot of information given about a topic, analyzing if the original statement is true or not is still a challenging task.
Simple questions Simple questions require small piece of text to find an answer, and thus, they are easier to comprehend. For instance, for a question like "*What is the magnitude of earthquake in Pakistan?*," it can easily be deduced that the answer of this question would be a simple numeric value. The process of finding an answer to a simple question consists of three basic steps: (i) question analysis, (ii) relevant documents/knowledge graphs retrieval, and (iii) answer extraction [7]. MS MARCO [57], SQuAD [75], and FreebaseQA [33] are some of the examples of simple question-based datasets.
Complex questions Complex questions are questions that require different types of knowledge or several steps to answer. They are difficult to answer and require access to multiple documents or multiple interactions with the system [4]. Complex question like "*how many cities in China have more population than New Delhi?*" requires the system to first figure out the population of New Delhi and then compare it with the population of different cities in China. Thus, answering complex questions requires complex techniques such as iterative query generation [66], multi-hop reasoning [107], decomposition into subquestions [32], and combining cues from the multiple documents [44]. Large-scale complex question answering dataset (LC-QuAD) [98] and Complex Sequential Question Answering (CSQA) [82] are some of the examples of complex QA datasets.
Casual questions Casual questions require detailed explanation pertinent to the entity, and they usually start with the words like *why* or *how*. The answers generated for casual questions are not straightforward or concise. This generation of detailed answers call for advanced natural language processing techniques that are able to understand the question on different levels of technicality such as semantics and syntax [27]. An example of such questions could be "*why do earthquakes occur?*."
Listing questions These are the questions which require the list of entities or facts as an answer, e.g., "*list the name of all the former presidents of America*." The techniques that are utilized to answer factoid question works well for the listing questions. The reason being that QA systems treat such questions as a sequence of factoid questions asked iteratively [52].
Unanswerable questions These are the questions whose answers cannot be found or deduced via the source text. Unanswerable questions could be any type of the aforementioned questions. For these questions, the correct result of the QA system is to indicate that it is unanswerable. SQuADRUn [76] is an extension of the SQuAD dataset [75] with over 50,000 unanswerable questions that was introduced to further improve the task of QA.

2.1.3 Types of data sources

CQA systems can be classified on the basis of the underlying data sources they utilize to find an answer. These underlying data sources could be:
Structured data source In a structured document, data is stored in the form of entities. These entities form a separate table. An entity in a table can have multiple attributes associated with it. The definition of these attributes is referred to as the metadata and is stored in a schema. A query language is used to access the data and retrieve relevant information from the schema. Examples of structured data sources are databases and Resource Description Framework

(RDF) graphs. Question Answering over Linked Data (QALD)[15] and LC-QuAD [98] utilize structured data source (i.e., RDF graphs) to answer the questions.

Semi-structured data source There is no clearly defined boundary between the stored data and its schema in the semi-structured data sources which makes it quite labor-intensive to build. An example of a semi-structured data source is XML. The datasets that are designed using semi-structured data sources include TabMCQ [102] and Question Answering using Semi-structured Metadata (QuaSM) [64].

Unstructured data source There are no pre-defined rules for storing the data in this particular arrangement. The data stored in the unstructured data sources could be of any type and require the use of advanced natural language processing techniques and information retrieval methods to find out the relevant answer. However, the reliability of finding the correct answers is low as compared to the structured data sources. Examples of unstructured datasets are SQuAD [75], Question Answering in Context (QuAC) [13], and CNN & Daily Mail [26].

2.1.4 Types of CQA systems

Over the past few years, the demand for CQA systems, from both research and commercial perspective, has increased in turn enabling users to search a large-scale knowledge base (KB) or a text-based corpora written in natural language. This categorizes the CQA systems into sequential KB-QA agents and CMRC:

Sequential KB-QA KB-QA systems are extremely flexible and easy to use in contrast to the traditional SQL-based systems that require users to formulate complex SQL queries [17]. In a real-world scenario, users do not always ask simple questions [82]. Usually, the questions asked are complex in nature and, therefore, require multi-turn interaction with the KB. Also, once a question has been answered, the user tends to put forward another question that is linked to the previous question–answer pair. This forms the task of sequential QA using knowledge graphs.

Conversational machine reading comprehension The practical use of text-based QA agents, also referred to as CMRC agents, is more common in the mobile phones than in the search engines (like Google, Bing, and Baidu), wherein concise and direct answers are provided to the users rather than presenting them with a list of possible answers. For instance, if a user intends to look for a popular restaurant in a particular geographical area, the search engine would provide her with a search result encompassing options spread on multiple pages, whereas a CMRC-based dialog agent would ask a few follow-up questions to figure out the preference(s) of the user to subsequently narrow down the search result to one, i.e., possibly the best, answer. With the emergence of CMRC, many researchers [13, 32, 77, 82] have tried inducing a conversational aspect to meet the requirements for the task of CQA by introducing a background context and a series of interrelated questions.

2.2 What makes CQA different from QA?

2.2.1 Task-based differences

The task of CQA differs from the traditional QA in a number of ways. In traditional QA systems, questions are independent of each other and are based on the given passage. In

[15] http://qald.aksw.org/

Table 1 A chunk of a dialog from the CoQA dataset [77]

Topic: Staten Island	
Passage:	Staten Island is one of the five boroughs of New York City in the US state of New York. In the southwest of the city, Staten Island is the southernmost part of both the city and state of New York, with Conference House Park at the southern tip of the island and the state. The borough is separated from New Jersey by the Arthur Kill and the Kill Van Kull, and from the rest of New York by New York Bay. With a 2016 Census-estimated population of 476,015, Staten Island is the least populated of the boroughs but is the third-largest in area at. Staten Island is the only borough of New York with a non-Hispanic White majority. The borough is coextensive with Richmond County, and until 1975 was the Borough of Richmond. Its flag was later changed to reflect this. Staten Island has been sometimes called "the forgotten borough" by inhabitants who feel neglected by the city government.
Question 1:	How many boroughs are there?
Answer 1:	Five
Question 2:	In what city?
Answer 2:	New York City
Question 3:	And state?
Answer 3:	New York
Question 4:	Is Staten island one?
Answer 4:	Yes
Question 5:	Where is it?
Answer 5:	In the southwest of the city
Question 6:	What is it sometimes called?
Answer 6:	The forgotten borough
Question 7:	Why?
Answer 7:	Because the inhabitants feel neglected by the city government

contrast, questions in CQA are related to each other which poses an entirely different set of challenges including but not limited to:

- In order to find the correct answer for the question at hand, the model needs to encode not only the current question and source paragraph, but also the previous history turns. More specifically, as shown in Table 1, Question 2 and Question 3 are related to Question 1.
- The turns in CQA are of different nature. Some questions require more detailed information (i.e., *drilling down*), some may require information about some topic previously discussed (i.e., *topic shift*), some may ask about a topic again after it had been discussed (i.e., *topic return*), and some questions may ask for the clarification of topic (i.e., *clarification question*) [110]. All of these characteristics are incremental in nature and present challenges that most of the top-performing QA models fail to address directly, such as pragmatic reasoning and referring back to the previous context applying co-reference resolution. In Table 1, Question 2 is an example of a drill down question, Question 7

is a clarification question and "*it*" in Question 5 "*where is it?*" requires co-reference resolution.

2.2.2 Architectural differences

The architecture of a CQA model is similar to the one of a QA system on the base level. However, to introduce the conversational touch to the system, a CQA model extends the traditional QA system by introducing a few modules:

- A traditional single-turn KB-QA system encompasses a semantic parser and a knowledge base reasoning (KBR) engine. In addition to these, a sequential KB-QA system encompasses a dialog manager, which is responsible for tracking the previous dialog states and determines what question to ask next to help a user search the KB effectively for an answer.
- A CMRC system differs from a traditional MRC system in two aspects. First, the encoder is embedded with a submodule referred to as history modeling module, which is responsible for not only encoding the current question and the given passage, but also the history turns of the conversation. Second, a reasoning module is extended to generate an answer, that might not be directly given in the passage, using pragmatic reasoning [77].

It is worth noting here that the paradigm of CQA is an emerging one, which is not well studied in contrast to traditional QA systems. Therefore, not many research papers are available. The architecture and researches carried out in both sequential KB-QA systems and CMRC systems will be discussed in detail in Sects. 3 and 5.

3 Sequential KB-QA systems

A knowledge base (KB) is a structured information repository used for knowledge sharing and management purposes [47]. Freebase [5], NELL [53], DBpedia [39], and Wikidata[16] are well-known examples of large-scale graph-structured knowledge bases also termed as the knowledge graphs (KGs) and have become significant resources when dealing with open-domain questions. The KGs are known to be a graphical representations of a KB, and a typical KG comprises of triples encompassing subject, predicate, object triples *(s,r,t)*, wherein *r* is a relation or predicate between the entities *s* and *t* [22]. They play an important role in bridging up the lexical gap by providing additional information about relations which in turn helps in gaining more detailed information about the context. The knowledge graphs have seen their successful applications in various NLP tasks such as text entailment, information retrieval, and QA [119]. The task of QA over large-scale KB-QA systems has seen its progress from simple single-fact task to complex queries requiring multi-hop interaction and traversal of the knowledge graphs. These come under the category of single-turn QA where a user puts forward a question and the system finds the best possible answer for it. Though KB-QA-based agents improved the flexibility of QA process to a considerable extent, nevertheless, it is irrational to believe that these systems could constitute complex queries without having complete knowledge about the organizational structure of the KB to be questioned [23]. Thus, sequential KB-QA system is a more optimal option as it lets the users query the KB interactively.

[16] https://www.wikidata.org/wiki/Wikidata:Main_Page.

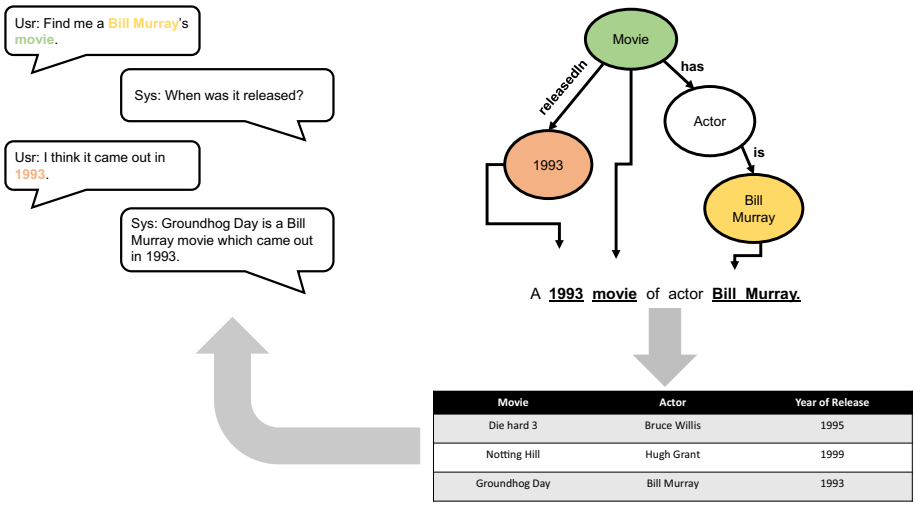


Fig. 5 Aligning knowledge and conversation in sequential KB-QA

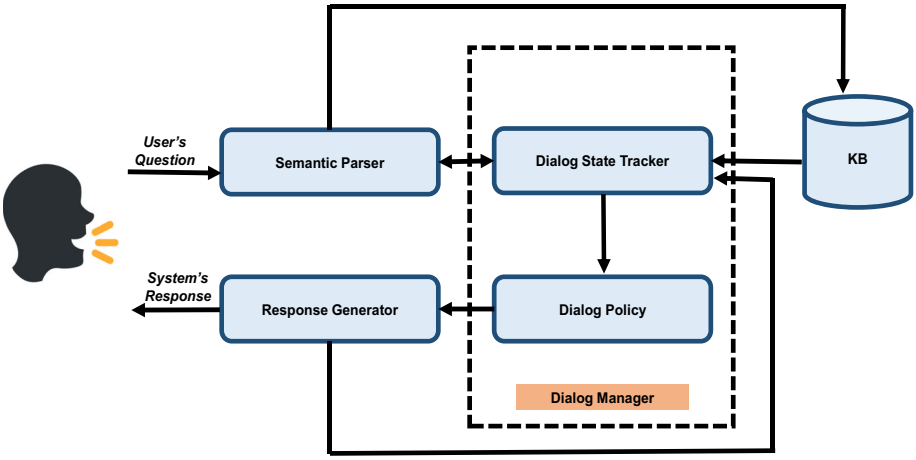


Fig. 6 A high-level diagram of sequential KB-QA

The interactive sequential KB-QA system is useful in many commercial areas such as making a restaurant reservation [93], finding a hotel in a new city, finding a movie-on-demand [19], or asking for relevant information based on certain attributes. Figure 5 illustrates how a sequential KB-QA system aims to find a movie based on specified attributes by a user. If it is a traditional KB-QA system, the conversation would have ended after the first turn with a number of results. But under the sequential KB-QA setting, the system asks the follow-up questions for the specific details about the current question and present the user with the most appropriate answer.

The core architecture of a sequential KB-QA system comprises of a semantic parser and an inference engine, along with the addition of a dialog manager, that keeps track of the previous turns and decides which questions to ask to help the user query the KB effectively. The high-level architecture of a sequential KB-QA is depicted in Fig. 6, which consists of: (i)

Semantic Parser, (ii) *Dialog Manager*, and (iii) *Response Generator*. The semantic parser is responsible for mapping input along with the previous context into a semantic representation (logical form) to query the KB. The dialog manager keeps track of the dialog history (i.e., QA pairs and DB state) and updates it accordingly [92]. It is also responsible for selecting the system's next action (i.e., to provide an answer or to ask a clarification question) based on the current question using dialog policy. The process of dialog policy can be either trained on dialogs [19, 104] or programmed [105]. At the end, the response generator converts the system's action into natural language response. However, certain new approaches [14, 55] are working toward the elimination of the semantic parser module as it requires extensive and expensive labeling of data.

3.1 Semantic parser

The notion of semantic parsing can be thought of as a process of mapping natural language text into meaningful logical forms and has emerged as a significant technical component for designing KB-QA systems [12]. Once a correct logical form has been obtained, it can be executed on the knowledge source in the form of a query to obtain answer denotations.

Iyyer et al. [32] introduced the task of semantic parsing for sequential QA by creating a dataset of simple interrelated questions out of a complicated WikiTableQuestions dataset [60]. The proposed model, called Dynamic Neural Semantic Parsing (DynSP), is a weakly supervised structured output learning approach based on the reward-guided search. Given a question along with a table, the model forms a semantic parsing problem as a state-action search problem wherein each state denotes a partial or complete parse and each action can be considered as an operation to extend the parse. Unlike traditional parsers, DynSP explores and constructs different neural network structures for different questions.

The aforementioned approach maps only the current utterance into its logical form which makes it difficult for the system to interpret the meaning of the utterance especially where co-reference resolution is required. To address this shortcoming, the authors in [23] proposed dialog-to-action (D2A) to facilitate the use of previous utterances (both questions and answers) concatenated with the current question. The task of generating logical form can be regarded as the prediction of a series of actions, and each of them corresponds to simple grammar rules.

However, the model of D2A suffers from the problem of error propagation as it learns to reproduce previously generated actions, which might be incorrect. To overcome the issue of error propagation and ambiguous entity linking, the stepwise framework is improved by multitask learning for sequential KB-QA systems [90]. This model, multitask Semantic Parsing (MaSP), learns pointer-based semantic parsing and entity detection simultaneously as they are closely related. The joint learning could enhance the performance of the CQA task. Specifically, the input consists of the current question and historical interactions are passed through an encoder based on Transformer [99] to generate the context-aware embeddings. The model employs pointer network [101] to locate the targeted entity and a number in the given question. The use of the pointer network comes with two advantages: (i) It handles the co-reference resolution by learning the context of the entity, and (ii) it reduces the size of decoding vocabulary significantly from several millions to several dozens. The model also incorporates a type-aware entity detection module in which the prediction is fulfilled in joint space of IOB (inside, outside, beginning) tagging and corresponding entity type for disambiguation. In the end, grammar-guided decoder is used to infer logical forms that can be executed on the KB.

The model of MaSP suffers from the issue of producing ambiguous results because the task of jointly learning predicate and entity classification share no common information except for the supervision signals propagated to the classifiers. Also, the model depends entirely on the context to locate the answers. The issue was overcome by another recently introduced model called **C**ontext tr**A**nsforme**R** s**T**acked p**O**inter **N**etworks (CARTON)[65] which propagate signals in sequential order, and all the components use the signal forwarded from the previous components. CARTON's stacked pointer networks incorporate knowledge graph information for performing any reasoning and does not rely only on the conversational context. Moreover, pointer networks provide the flexibility for handling out of vocabulary which was not supported by MaSP. CARTON also proposed a new semantic grammar over MaSP with new logical rules. These rules helped the model to improve the overall performance. Another model, called mu**L**ti-task sem**A**ntic par**S**ing with tr**A**nsformer and **G**raph atte**N**tion n**E**tworks (LASAGNE) [35], goes a step beyond to improve MaSP. The model performs multitask learning by utilizing a Transformer [99] supplemented with a graph attention network (GAT)[100]. The model uses a Transformer to generate the logical forms of a natural language question, while GAT model is utilized to exploit the correlations between predicate and entity types due to its message-passing ability between the nodes. The authors also proposed an entity recognition module that contributes in detecting, linking, filtering, and permuting all the relevant entities in the generated logical forms. Unlike MaSP and CARTON, LASAGNE use both sources of information, the encoder and the entity recognition module to perform these operations which makes the process of re-learning entity information from the context of the current question avoidable.

3.2 Dialog manager

Conversational history plays a significant role when generating the logical forms of natural language utterances. Once a logical form is obtained, the system is in a better state to decide its next action, i.e., to ask a clarification question or provide an answer to a question.

Dialog-to-action [23] incorporates a dialog memory to store the historical interaction of a user. The model consists of a bidirectional RNN with a gated recurrent unit (GRU) [15] as an encoder to convert the input (previous question–answer pairs concatenated with the current question) into a sequence of context vector. A grammar-guided decoder (GRU with attention mechanism) generates an action sequence based on the context vector [45]. The dialog memory used in the model encompasses entities, predicates, and action subsequences which could be replicated selectively as decoding proceeds.

Both CARTON [65] and LASAGNE [35] incorporate the dialog history based on previous interactions as an additional input to the model for handling ellipsis and co-reference. The final input consists of the previous question–answer pair and the current question. The utterances are separated using a *[SEP]* token and at the end of the last utterance, a context token *[CTX]* is appended. The conversation is tokenized using WordPiece tokenization [106] and then pre-trained Global Vectors (GloVe) model [62] is used to embed the words into vector representations.

3.3 Response generator

In NLP tasks, response generation is the last and vital step to generate system utterances for a user, and the introduction of pre-trained language models has been a game-changing factor for the promising field of language generation over the past few years. Peng et al. [61]

introduced a model based on Open AI's Generative Pre-training (GPT) [72] called semantically conditioned generative pre-training (SC-GPT). The paper introduces a dataset called FewshotWOZ[17] to simulate the process of few-shot learning for limited data labels. SC-GPT generates semantically controlled responses and is trained in three steps: (i) Initially, it is pre-trained on massive plain corpora so that it can better generalize to new domains, (ii) further pre-training is conducted on dialog-act-specific huge corpora to gain the capability of controllable generation and finally, and (iii) a limited amount of domain labels is used to fine-tune the model for its adaptation to the target domain.

Another framework called NLG-LM [118] employs multitask learning to not only generate semantically correct responses, but also maintain the naturalness of the conversation. The model utilizes sequence-to-sequence architecture to simultaneously train the natural language generation (NLG) and language modeling (LM) tasks. The language modeling task, carried out in decoder, is incorporated on human-generated utterances to bring out more language-related elements. In addition to that, the unsupervised nature of the language model eliminates the need for a massive amount of unlabeled data for training purposes.

3.4 Sequential KB-QA approaches without semantic parser

There exists extensive research work in semantic parsing, wherein deep neural networks have been utilized for training models in a supervised learning setup over manually generated logical forms. However, generating labeled data for this task could be exhaustive and expensive [55]. To address this issue, a new research direction has been recently investigated that utilizes weak training for semantic parsing where training data consists of question and answers and the structured resources are used to restore the logical representations that would result in the right answer.

In [82], the authors proposed a model which is an amalgamation between hierarchical recurrent encoder–decoder (HRED) [87] model and key-value memory network [50] to present the fusion of dialog and QA process. HRED is responsible for generating high-level and low-level representations of an utterance and the context. Candidate tuples are selected in which the entity appears as subject/object. These candidates, i.e., tuples are stored in a key-value memory network as key-value pair, where the key contains the relation–subject pair and the value contains the embeddings of the object. The model makes multiple passes (turns) to attend to different aspects of the question especially in the case of complex questions. A decoder is used to generate answer sequences.

Another approach in [55] presents a table-centered sequential KB-QA model which, instead of learning the intermediate learning forms, encodes the structured resources (i.e., tables) along with the questions and answers from the conversational context. The approach encodes tables as graphs by representing cells, columns, and rows. The column represents the main features of the questions and cells contains the relevant values. To handle the follow-up questions, the model adds previous answers by marking all the columns, rows, and cells with nominal features. It uses a graph neural network (GNN) [84]-based encoder to encode the graph by generating vector representation of the edge label between the two nodes. The copy mechanism based on the pointer network, instead of selecting symbols from output vocabulary, then predicts the sequences of answer rows and columns from the given input.

CONVersational KB-QA with context **EX**pansion (CONVEX) [14] employs unsupervised method to answer sequential questions (follow-up questions) by keeping track of the conversational context using predicates and entities appeared so far. The initial question is

[17] https://github.com/pengbaolin/SC-GPT.

Table 2 Recent studies on sequential KB-QA (2016-2021)

Refs.	Contribution(s)	Techniques used	Merits	Demerits
DynSP [32]	Introduced the task of semantic parsing for sequential KB-QA.	Dynamic Neural Network structure.	Reward-guided search reduces the number of queries to be labeled.	The parse language is not comprehensive enough to represent the semantic parses of the sentences in dataset.
				The table-based search-space approach cannot be scaled up to cater the needs of large-sacle curated KGs.
HRED + KVmem [82]	Introduced sequential KB-QA dataset consisting of complex questions.	End-to-end model based on HRED and KV memnet.	Incorporates dialog history with the current utterance.	Performs poorly with complex questions.
			Works well with simple and direct questions.	Doesn't work well with indirect or incomplete questions.
				KV memnet has flat organization of story which makes it unsuitable for complex questions.
D2A [23]	Introduced history interaction as a part of input to deal with enormous ellipsis phenomena.	A bidirectional RNN with a GRU is used as an encoder. A grammar-guided decoder along with a dialog memory component is used to generate action sequences.	The model can effectively handle the contextual references.	Error propagation may occur because the model replicates previously generated action sequences which might be incorrect.
			The parser introduced is capable of parsing various types of question.	The supervision signals cannot be shared among the model for mutual benefits as they are learned independently for the subtasks.
				Ambiguous entity linking

Table 2 continued

Refs.	Contribution(s)	Techniques used	Merits	Demerits
MaSP [90]	Multitask learning for sequential KB-QA.	Utilizes Transformer as a contextual encoder and a pointer-equipped decoder.	Reduces the risk of error propagation by jointly learning semantic parsing and entity detection.	May result in spurious logical form.
			Works well with co-reference resolution.	
			Addresses ambiguous entity linking by leveraging contextual features of the input.	
CARTON [65]	Improved multitask semantic parsing by introducing new logical grammar rules.	Utilizes Transformer model to generate logical forms, while the three stacked pointer networks are used for completing the final executable logical form against the KG.	Introduces the new set of vocabulary with advanced logical rules which results in improved model's performance.	May result in spurious logical forms.
			Works well with co-reference resolution.	There is no feedback signal from the resulting answer generated from its logical structure which hinders the model's learning in generating correct logical form.
			Improves the process of entity detection and linking by utilizing information from both entity detection module and encoder.	

Table 2 continued

Refs.	Contribution(s)	Techniques used	Merits	Demerits
LASAGNE [35]	Improved multitask semantic parsing for sequential KB-QA.	Utilizes Transformer model to generate logical forms, while the graph attention model is used to exploit correlations between entity type and predicates.	Eliminates the risk of producing ambiguous results by sharing signals between entity and predicate nodes.	May result in spurious logical forms which affects the model's performance in answering clarification and ellipsis-based questions.
		Introduced an entity detection module which detects, links, and permutes all the relevant entities.	Works well with co-reference resolution.	
			Improves the process of entity detection and linking by utilizing information from both entity detection module and encoder.	
GNN + PointerNet [55]	Conversation processing around structured data.	Neural approach based on GNN and pointer network.	Eliminates the need of semantic parsing.	Table-search methods cannot scale to large real-world KGs involving qualitative or logical comparison.
			Handles conversational context stored in tables, effectively.	
CONVEX [14]	Completion of incomplete follow-up questions.	Symbolic approach.	Automatically infers missing or ambiguous pieces for follow-up questions.	May result in combinatorial explosion if subgraphs not expended carefully
			Eliminates the need for intermediary representation of the context and given question.	

used for initializing and selecting a small subgraph of the knowledge graph. The essence of this approach is the graph exploration algorithm that tends to expand a frontier aptly to find the possible candidate answers for the follow-up questions. The right answer from the candidate answers is selected by calculating weighted proximity. The top-scoring answer (in the range of 0 to 1) will be returned as the answer to the current question.

Table 2 summarizes the major contributions, the techniques exploited, and the merits and demerits of the aforementioned approaches.

4 Conversational machine reading comprehension

Most of the work carried out in the field of MRC is based on single-turn QA which is unlikely in the real-world scenario since humans tend to seek information in a conversational context [78]. For instance, a user might ask, "*Who is Christopher Columbus?*" and based on the answer received, he might further investigate, "*Where was he born?*" and "*What was he famous for?*." It is easy for a human to decipher that here "*he*" in the follow-up questions refer to "*Christopher Columbus*" from the first question. But when it comes to a machine to comprehend the context, it poses a set of challenges such as co-reference resolution or conversational history [42], which most of the state-of-the-art QA systems do not address directly.

A typical MRC model consists of three main functions, namely (i) encoding the given context and question into a set of symbolic representations called embeddings in a neural space, (ii) reasoning through the embeddings to find out the answer vector in the neural space, and (iii) decoding the answer vector to produce natural language output [22]. In [68], the authors proposed a modification by introducing two modules, i.e., history selection and history modeling modules to address the aforementioned challenges to incorporate the conversational aspect, hence introducing the task of CMRC. Formally, given a context C, the conversation history in the form of question–answer pairs $Q_1, A_1, Q_2, A_2, \ldots, Q_{i-1}, A_{i-1}$, and a question Q_i, the CMRC model needs to predict the answer A_i. The answer A_i can either be a free-form text with evidence [77] or a text span [13]. The flow of a general CMRC model is depicted in Fig. 7.

We will discuss these modules separately in the rest of this section, along with the techniques and trends utilized in each of them for the successful design and implementation of a CMRC model.

4.1 History selection module

To enable the CMRC model to predict the answer span more accurately, it is necessary to introduce the previous context along with the source passage and current question. However, context utterances that are relevant to the query are useful, whereas the irrelevant ones may bring more noise [96, 113]. Thus, the careful selection of conversational history turns is quite critical for the model. History selection process can be categorized as:

4.1.1 Selecting K turns

Contextual attention-based deep neural network (SDNet) [117], bidirectional attention flow (BIDAF++) [13], open-retrieval CQA (ORConvQA) [70], and weakly supervised open-

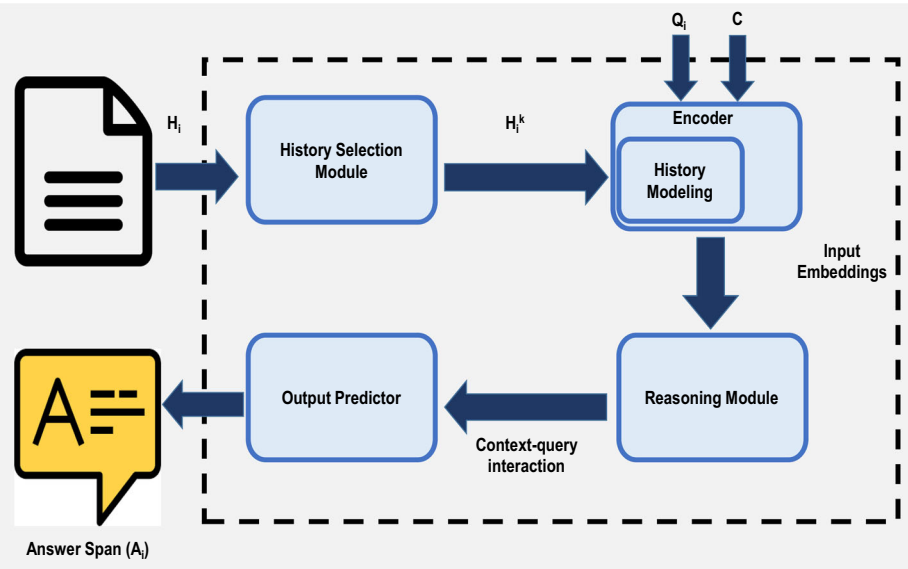


Fig. 7 Generic framework of a CMRC model which consists of (i) history selection module that selects $H_i{}'$ history turns from the conversational history context H_i, (ii) encoder that transforms the tokens of $H_i{}'$, C, Q_i into input embeddings, (iii) reasoning module is responsible for performing contextual integration of input embeddings into contextualized embeddings to perform reasoning, and (iv) output predictor predicts the answer A_i on the basis of context-query interaction

retrieval CQA (WS-ORConvQA) [71] utilize conversation history by incorporating K rounds of history turns.

4.1.2 Immediate history turns

BERT with 2-ctx [58] suggests that incorporating immediate two turns can be helpful in predicting the right answer span, whereas BERT-HAE [68] claims that incorporating 5–6 conversational history turns contributes more in finding the correct answer span. However, both models demonstrate a dramatic degradation in the performance with the increase in the number of history turns.

4.1.3 Dynamic history selection

In [110], the authors pointed out that the dialog features like topic return or topic shift may not align with the concept of selecting immediate dialog turns. Therefore, in order to address this shortcoming, History Answer Modeling (HAM) [69] was introduced as a dynamic policy that weighs the previous dialog turns on the basis of their contribution to answering the current question. The model assigns weight by attending the previous history turns at a token level or sentence level and combine the same with the current turn's representation.

Another approach, referred to as Env-ConvQA [67], proposed a dynamic k-history turns selection process based on reward-based reinforced backtracking policy. The model treats the process of extracting the relevant history turns as a sequential decision making process. The model acts on the provided history turns and backtracks through each turn one by one to decide whether the turn is relevant to the current question or not.

4.2 Encoder

This component is responsible for converting the tokens of the source passage, current question, and the selected history turns into fixed-length vectors which are subsequently provided as an input to the reasoning module. Although the internals of an encoder may vary from approach to approach depending on the input required by the reasoning module, nevertheless, the high-level encoding generally involves transforming and combining different context-dependent word embeddings, including but not limited to, Embeddings from Language Model (ELMo) [63], GloVE [62], and BERT [18]. To improve the impact of these embeddings, additional features such as Parts of Speech (POS) tags and History Answer Embeddings (HAE) have also been incorporated as a part of the input. These embeddings can be categorized into conventional word embeddings and contextualized word embeddings.

4.2.1 Conventional word embeddings

This technique is responsible for encoding of words into low-dimensional vectors. The encoding is done in such a way that the interrelated tokens are placed in close proximity to each other in vector space to make the identification of co-relation easy between them. Several methods for generating distributed word representation have been proposed in the literature, with the most popular and efficient being GloVE [62] and Word2Vec [49]. However, these methods fail to determine the accurate meaning of the words with respect to their given context.

4.2.2 Pre-trained contextualized word embeddings

Though the conventional word embeddings method yields good results in identifying and establishing the correlation between the words encoded in low-dimensional vectors, they still fail to capture the contextual representations sufficiently. To be accurate, the distributed word representations generated for a single word are the same in varying contexts. To overcome this issue, the idea of contextualized embeddings was put forward by the researchers. These embeddings are pre-trained on large corpora of text and are then utilized as either distributed word embeddings or are fine-tuned according to the specific downstream task. This comes under the category of transfer learning and has obtained astonishing results in various NLP-based tasks [18, 63, 109].

The most successful application of these embeddings has been in the field of machine comprehension. One of the very first in the series is Context Vectors (CoVe) [48] which utilizes Seq2Seq models [95] to train long short-term memory (LSTM) [28] encoders on a large-scale dataset. The encoder then utilizes the obtained results on other downstream NLP tasks. Proposed by [63], ELMo is a successor of CoVE and embeddings are obtained by training a bidirectional language model (biLM). These embeddings can generate more accurate representations of the words as instead of using the results from the topmost layer of biLM, it combines outcomes from all the layers of biLM into one vector and assign a weighting score that is task-specific. Another popular model in terms of language understanding is Transformer [99] which is a sequence transduction model based on multi-headed attention, thus entirely eliminating the need of utilizing multiple recurrent layers that are part of the most encoder–decoder architectures. This mechanism of self-attention makes the Transformer model more efficient and parallelizable in learning the context of the input sequences. The most recent and top-trending one in the series is BERT [18] that has addressed the issue

of unidirectionality used in training of different language models such as Generative Pre-training (GPT) [72] and GPT-2 [73]. Due to the bidirectional property and the powerful Transformer [99] architecture, BERT's performance exceeds the top-performing models in many NLP downstream tasks [18].

4.3 History modeling

The process of history modeling is generally carried out in the encoder module where the conversation history is integrated with the context and current question to form a complete input.

We describe it as a separate module for easy understanding and better readability. Different models employ different techniques or a combination of these to introduce conversational history turns as a part of an input. A brief description of each of these techniques is given as follows:

4.3.1 Appending the conversation history

One of the most common ways to include the selected history turns (previous question–answer pairs) as a part of the input is by appending them with the current question [67, 70, 117] . This is further modified at sublevel by some approaches [11, 58] via appending only history questions along with the turn number encoded with it. In [13], the authors claimed that adding dialog turn in the input yields better results practically.

4.3.2 Introducing history answer markers in the given context

Another trend seen recently in modeling the conversation history is encoding the context tokens in history answer embeddings markers [68]. The advantage of using these tokens is that they work as an indicator to point out whether a context token is a part of history answer or not. Another variation of HAE is Positional HAE (POS-HAE) [69], wherein position information of dialog turn relative to the current question is also encoded. This enables the model to capture the spatial patterns of history answers in context.

4.3.3 Generating latent representations using context tokens

One of the attributes of successful CMRC models is being able to grasp the flow of the conversation. Since the flow of the conversation based on the given context, it can be captured by generating latent or intermediate representations of the context tokens rather than using the raw inputs. Such approaches [29, 111] fall under the category of flow-based methods.

4.4 Reasoning module

CMRC models can be grouped based on how they perform the process of reasoning. For *single-step reasoning*, the model passes the contextualized input (context, question, and history turns) only across one layer and generates the answer. In contrast, for *multi-step reasoning*, the contextualized input is fused across multiple layers to produce history-aware contextualized output embeddings. Generally, the input for this module consists of multiple sequence sets which are then fused in multiple layers and are usually interwined with an

attention mechanism to generate accurate output embeddings. On the basis of underlying techniques, the reasoning process can be categorized as *conventional methods*, *pre-trained language models*, *flow-based models*, and *open-retrieval-based models*.

4.4.1 Conventional methods

Several sequence models employing different mechanisms like self-attention and bidirectional attention are a common choice for carrying out the task of CMRC. Famous as CoQA's baseline, DrQA+PGNet [77] leverages the strengths of two powerful models, i.e., Pointer-Generator Network (PGNet) [85] and Document Reader (DrQA) [10]. DrQA, based on bidirectional LSTM (biLSTM), first provides cues from the answer evidence in the given context. PGNet, which utilizes an attention-based Seq2Seq model [1], decodes the found evidence to predict the final answer.

BiDAF++ [13] uses the bidirectional attention flow (BiDAF) [86] model augmenting the bidirectional attention flow along with contextualized embeddings and self-attention. The modeling performs reasoning via a multilayered bidirectional attention flow layer followed by a multilayered biLSTM to identify the correct answer span. SDNet [117] utilizes two bidirectional Recurrent Neural Networks (RNNs) [80] to apply both self-attention and inter-attention between different layers in order to form the contextualized understanding of question and context.

4.4.2 Pre-trained language models

Large-scale pre-trained language models such as BERT [18], RoBERTa [43], and GPT [72] have become popular to achieve the state-of-the-art results on NLP tasks. While GPT is known for its language generation capabilities, BERT is famous for language understanding and has provided great results in machine comprehension tasks. One of the advantages of employing pre-trained language models is their capability to fuse both encoding and reasoning modules together. This results in a ready-to-tune architecture that hides the complex interactional nature between the given context and current question. However, incorporating previous context is a challenging task in pre-trained language models (particularly BERT) as it allows for only two segments in the input and the length of sequence is limited to 512. The more turns we try to append, the more context paragraph or history turns need to be truncated to be able to adapt to the model. The accurate modeling of the history results in better reasoning over the context. The history integration challenge can be addressed using the following approaches:

- Highlighting conversational history by embedding history answer embeddings in the contextual tokens as suggested in BERT-HAE [69]. The embeddings are only added for those tokens that are present in the previous conversational history.
- Using separate models for all the history turns to attend to the interaction between each turn and the given context as suggested by Ohsugi et al. [58]. The contextualized embeddings are then merged together to form an aggregated history-aware embeddings. These aggregated embeddings are then passed from BiGRU to capture an inter-turn interaction before any prediction can be made.
- Introducing a reinforced backtracker in the model to filter out the unnecessary or irrelevant history turns instead of evaluating them as a whole as proposed by Qiu et al. [67]. The selected turns along with the given passage forms an input to be provided to the BERT model.

Once the history turns have been integrated, BERT-based models calculate the probability of each word being the start word by generating a dot product between the final embedding and the start vector, followed by the application of softmax over all the words [68]. Finally, the word with the highest probability value is selected. A similar process is employed to locate the final word in the given context. In [67], the model after predicting the answer span generates a reward to evaluate the utility of the history selection for answer prediction process. The computed reward, in turn, is utilized to update the policy network to maximize the accuracy of the model for the next cycle of prediction.

4.4.3 Flow-based models

Another recent trend that has caught attention is the use of flow-based approaches in machine comprehension. A well-designed CMRC model should be able to grasp the flow of the conversation, i.e., knowing what topic is under discussion as well as facts and events relevant to it. Thus, the flow of conversation can be considered as a sequence of latent representations generated based on the token of source passage. These latent representations, generated during the reasoning of previous conversations, aid in the contextual reasoning of the current question. The main models based on flow architecture are described below.

FlowQA [29] utilizes the contextualized embeddings as the latent representations, a process often referred to as Integration Flow (IF). The process involves the sequential processing of the context tokens in parallel to the question turns (referred to as context integration) along with processing question turns sequentially parallel to context tokens (flow). The model utilizes multiple flow layers interweaved with attention first on the context and then on the question itself to come up with the reasoning for answer span. FlowDelta [111] was introduced as an improved version in the flow series that utilizes the same architecture as FlowQA but achieves better accuracy. Instead of using the intermediate or latent representations, the model passes the information gain through the reasoning process. The information gain is nothing but the difference between the latent representations of the previous two layers. By modeling such difference, the model would better focus on the information hints present in the context.

The previously discussed flow approaches follow the concept of IF that does not really mimic a human's style of reasoning. The underlying reason is that they first perform reasoning in parallel for each question and then refine and enhance the reasoning across different turns. Graph Flow [11], on the other hand, constructs a dynamic context graph encoding not only the passage itself but also the question as well as the conversation history. The model processes the flow by applying GNN on all the sequences of context graphs and the output is utilized when processing the next graph. To capture the contextual relationship between the words, a biLSTM is applied before providing the words as an input to GNN. The Graph Flow architecture alternates this mechanism with co-attention over the question and the GNN output.

4.4.4 Open-retrieval-based models

Another recently introduced trend in the field of CMRC is the use of open-retrieval methods. The methods discussed above relies heavily on the given passage to extract or generate an answer. However, this seems impractical in real-world scenario since the availability of gold passage is not always possible. Thus, the model should be able to retrieve the relevant passages from a collection. The main models employing the open-retrieval architecture are discussed below:

ORConvQA [70] is first in the series of open-retrieval models for CMRC. It consists of three main modules: (i) a passage retriever, (ii) a passage re-ranker, and (iii) a passage reader. The three modules are based on Transformers [99]. The passage retriever first extracts the top-K relevant paragraphs from a collection provided a current question and the previous history. The retriever is based on dual-encoder architecture that utilizes two separate ALBERT [38] encoders for passages and questions. The re-ranker and reader uses the same BERT encoder. The encoder transforms the input sequence consisting of question, history, and relevant passages into the contextualized representations to be utilized by re-ranker and reader for answer extraction. The re-ranker module conducts a list-wise re-ranking of the retrieved passages which serves as a supervision signal to fine-tune the encoder. In the end, answer span is predicted by the reader module by computing the probability of the tokens being a start/end token.

In ORConvQA, the model focuses on identifying and extracting short span-based answers. In information-seeking dialog, however, answers are relatively free-form and long which are difficult to extract. WS-ORConvQA [71] is an extension of ORConvQA and introduces a learned weak supervision approach that can find and extract both span-based and free-form answers. And if the exact match is not found, the model tries to find a span in the retrieved passages that has the maximum overlap with the gold answer. Given a question and its conversation history, the passage retriever first extracts the relevant paragraphs from a collection. The retriever assigns a score based on the dot product of the representations of the questions and the passage. The reader then reads the top passages and produces an answer. The model works on weakly supervised training approach. Given one of the retrieved passages and gold answer, the weak supervisor predicts a span in the passage as weak answer to provide weak supervision signals for training the reader. The reader is based on standard BERT-based machine comprehension model [18] that calculates the probability of tokens being a start and an end token. The final answer is selected by computing the sum of its retriever score and reader score.

4.5 Output prediction

The common trends that have been observed for the answer prediction module include span prediction, free-form answer prediction, and dialog acts prediction. For span prediction, the probabilities of tokens being the end and start token is calculated. For unanswerable questions, a token, UNANSWERED, is appended at the end of each passage in QuAC. The model learns to predict this token if it finds the question unanswerable. A sequence-level aggregated representation is used to calculate dialog-act prediction and the modeling of history dialog acts is not required for the prediction of this task.

The categorization of the architecture based on the techniques used in each module is summarized in Table 3.

5 Datasets for conversational question answering

One driver for the rapid growth in the field of CQA is the emergence of large-scale conversational datasets for both knowledge base and machine comprehension. Constructing a high-quality dataset is equally significant as optimizing CQA-based architectures. In this section, we collect and compare the major datasets in the area of CQA.

Table 3 Recent studies on conversational machine reading comprehension (2016-2021)

Refs.	History selection	Encoder	History modeling	Reasoning	Output prediction
BIDAF++ w/k-ctx [13]	*k* history turns	GloVE for word embeddings.	Encodes context tokens with history answer markers before passing on for reasoning.	Performs reasoning via multilayered bidirectional attention flow layer followed by multilayered biLSTM.	Span prediction.
		Bidirectional LSTM for contextual embeddings.	Encode dialog turn number within the question embeddings.		
DrQA +PGNet [77]	*k* history turns	Bidirectional LSTM.	Appends the selected history turns to the source passage and current question.	DrQA model first point toward the evidence in the given text, PGNet then transform the evidence into the answer.	Free-form answers.
SDNet [117]	*k* history turns	Word embeddings using GloVe.	Appends the selected history turns to the source passage and current question.	Utilizes both self-attention and inter-attention in multiple layers using bidirectional LSTM to reason across the given context.	Span prediction.
		Contextualized embeddings using BERT.			
BERT-HAE [68]	*k* history turns but found optimal answer in 5 and 6 history turns.	BERT-generated embeddings.	Introduce history answer marker layer to the context token is present in any conversational history answer or not.	BERT generates a representation for each token based on the embeddings for position, segment, and tokens.	Span prediction.
				The model then computes the probability of tokens in a given paragraph of being a start and end token of the answer span.	

Table 3 continued

Refs.	History selection	Encoder	History modeling	Reasoning	Output prediction
BERT-HAM [69]	Dynamic history selection policy.	Bert-based embeddings on both word and sequence level.	Encode context tokens with dialog turn encoded variant of HAE called *Positional HAE*.	History attention module assigns weight to each token-level and sequence-level representation. And then aggregated representations of both are obtained that are further used for answer prediction.	Span prediction.
					Dialog-act prediction.
BERT w/k-ctx [58]	k history turns	Contextualized paragraph representations independently conditioned with each question and each answer generated using BERT.	Appends history QA pair to the current question with each QA pair conditioned on the source paragraph.	The concatenated result is then passed through the BiGRU for span prediction.	Span prediction.
			The model then concatenates the resulting sequences to form a uniform representation.		Answer type prediction (Yes, no, unanswerable).

Table 3 continued

Refs.	History selection	Encoder	History modeling	Reasoning	Output prediction
Env-ConvQA [67]	dynamic k history turns.	BERT-generated embeddings.	Prepends selected subset of history QA pair and passage to the current question.	BERT generates a representation for each token based on the embeddings for position, segment, and tokens.	Span prediction.
			The model then concatenates the resulting sequences to form a uniform representation.	The model then computes the probability of tokens in a given paragraph of being a start and end token of the answer span.	
				After answer prediction, the model generates a reward to evaluate the role of selected history turns and update the policy network accordingly.	
FlowQA [29]	k history turns.	Uses ELMo to generate contextual embeddings before passing it to IF layer.	Integrates both QA pairs and the intermediate context representation from conversation history called **FLOW**.	Employ multiple integration flow layers with alternating cross- and self-attention to perform reasoning.	Span prediction.
Graph flow [11]	Prepends N question–answer pairs to the current question.	GloVE and 1024-dim BERT embeddings.	Encodes history QA pairs into contextual graphs.	BiLSTM is utilized for the context integration and the GNNs are used to capture the contextual interaction.	Span prediction.
FlowDelta [111]	k history turns	Uses ELMo to generate contextual embeddings before passing it to IF layer.	Integrates both QA pairs and the intermediate context representation from conversation history called **FLOW**.	Model passes the information gain (the difference between the latent representations of last two layers) to let the model focus more precisely on the context.	Span prediction.

Table 3 continued

Refs.	History selection	Encoder	History modeling	Reasoning	Output prediction
ORConvQA [70]	k history turns	• Uses ALBERT to generate contextual embeddings before passing it to reader and re-ranker modules.	• Appends history questions to the current question.	• Employs fully supervised setting for the training of the reader.	Span prediction.
			• The model uses two encoders, one for encoding current question with its history and other for encoding relevant passages.	• The top-retrieved passages are then fed to the re-ranker and reader for a concurrent learning of all model components.	
				• The reader predicts an answer by computing scores of each token being the start token and the end token.	
WS-ORConvQA [71]	k history turns	• Uses ALBERT to generate contextual embeddings before passing it to reader module.	• Appends history questions to the current question.	• Employs weakly supervised setting for the training of the reader.	Span prediction.
			• The model uses two encoders, one for encoding current question with its history and other for encoding relevant passages.	• The top-retrieved passages are then fed to the reader.	Free-form answer.
			• The retriever generates a score based on the dot product of the representations of the question and the passage.	• The reader computes the probabilities of the true start and end tokens among all the tokens from the top passages.	
				• The answer span is selected on the basis of sum of retriever score and reader score.	

5.1 Datasets for sequential KB-QA

Most of the datasets for sequential KB-QA deals with simple questions, wherein each of them can be answered using a single tuple in the knowledge graph. However, in practice, a system can encounter a more complicated form of questions requiring it to use logical and comparative reasoning to come up with an accurate answer. The point worth noting is that unlike the simple questions, the complicated questions require access to the larger subgraph of the KG. For example, to answer the question, "*Which country has the highest peak, Nepal or India?*," one needs to find (i) the highest peak in Nepal, (ii) the highest peak in India, and finally, (iii) the comparison of both the peaks to come up with the right answer.

Similar to the field of CMRC, sequential KB-QA saw its rise after the introduction of sequential QA datasets, namely *Sequential Question Answering (SQA)* [32], *Complex Sequential Question Answering (CSQA)* [82], and *ConvQuestions* [14]. These datasets have facilitated the process of answering complex questions, thus supporting a number of researches. A high-level comparison based on their common characteristics is presented in Table 4.

5.1.1 SQA

The main idea behind the creation of SQA is to decompose the complex questions and convert them into a series of interlinked sequential questions to give a touch of natural conversation.

Dataset collection As described in [32], the SQA dataset has been collected via crowdsourcing by leveraging WikiTable Questions (WTQ),[18] which contains highly compositional questions associated with HTML tables from Wikipedia. Each crowdsourcing task contains a long and complex question originally from WTQ as the question intent. The workers are asked to compose a sequence of simpler but interrelated questions that lead to the final intent. The answers to the simple questions are subsets of the cells in the table.

Dataset analysis SQA consists of 6,066 unique question sequences containing 17,553 question–answer pairs resulting in an average of 2.9 questions per sequence. The questions are identified into three different classes: (i) *column selection* questions, wherein the answer is the entire column of the table and constitutes 23% of the questions in SQA, (ii) *subset selection* questions where the answer is the subset of the previous question's answer and contributes 27% of the questions in the dataset, and (iii) *row selection* questions where answers to the questions appear in the same rows but in different columns, making 19% of the dataset.

Evaluation For the system to be evaluated, the overall accuracy, sequence accuracy (the percentage of sequences for which every question is answered correctly), and positional accuracy (accuracy at each position in a sequence) are calculated. With that said, all systems struggle to correctly answer all questions within a sequence, despite the fact that each question is simpler on average than those in WTQ.

5.1.2 CSQA

The CSQA dataset [82] consists of 200K QA dialogs for the task of complex sequential question answering. CSQA combines two subtasks: (i) answering factoid questions through

[18] https://github.com/ppasupat/WikiTableQuestions.

Table 4 A comparison of the sequential KB-QA datasets SQA[32], CSQA[82], and ConvQuestions[14] based on different characteristics as defined in their respective papers

Characteristics	SQA	CSQA	ConvQuestions
Data source	• WikiTableQuestions	• WikiData	• WikiData (consisting of 5 domains, i.e., books, movies, soccer, music, and tv series).
Conversational setup	• Three workers who were asked to decompose complex sentences into a sequence of simpler sequential sentences.	• Pair of in-house annotators where annotator acts as a *user* and the other as a *system* to provide answers or ask clarification questions.	• Master workers from AMT paired together where they were asked to provide answers vis web search.
Nature of QAs	Simple	Complex interrelated as well as simple.	Complex
Question types	Factoid	Factoid	Factoid and non-opinionated.
Requires reasoning?	No	Yes	Yes
Max turns per dialog	N/A	8.5	5
Total number of questions	15, 553	1.6 M	N/A
Total number of dialogs	6066	200K	11,200

complex reasoning over a large-scale KB, and (ii) learning to converse through a sequence of coherent QA pairs. CSQA calls for a sequential KB-QA agent that combines many technologies including (i) parsing complex natural language queries, (ii) using conversation context to resolve co-reference and ellipsis in user utterances like the belief tracker, (iii) asking for clarification questions for ambiguous queries, like the dialog manager, and (iv) retrieving relevant paths in the KB to answer questions.

Dataset collection Each dialog is prepared in a two-in-house-annotators setting, one being a *user* and the other acting as a *system*. A user's role is to ask questions and a system's job is to answer the questions or asks for clarification if required. The idea is to establish the understanding of the simple and complex questions that can be asked by the annotators over a knowledge graph. These could then be abstracted to templates and utilized to instantiate more queries involving different objects, subjects, and relations. Apart from asking and answering simple questions (that requires only a single tuple to generate an answer), the annotators come up with questions involving logical and comparative operators like AND, OR, NOT, ==, and >=, resulting in more complex questions to judge model's performance. The examples of such questions are "*Which country has more population than India?*," and "*Which cities of India and Pakistan have River Indus passing through them?*." After collecting both simple and complex questions, the next step is to create coherent conversations involving these QA pairs. The resulting conversation should have (i) linked subsequent QA pairs, and (ii) the conversation should contain the necessary elements of a conversation such as confirmation, clarification, and co-references.

Dataset analysis The dataset consists of 200K dialogs and a total of 1.6 million turns. On average, the length of a user's questions is 9.7 words and a system's response is based on 4.74 words.

Evaluation: Different evaluation metrics are used to evaluate the different question types. For example to measure the accuracy of simple questions (consisting of indirect questions, co-references, ellipsis), logical reasoning, and comparative reasoning, both precision and recall are used. When dealing with quantitative reasoning and verification (Boolean) questions, F1 score is utilized. For clarification questions, BLEU-4 score is used.

5.1.3 ConvQuestions

ConvQuestions [14] has been published recently to further aid the field of sequential KB-QA. It consists of 11,200 distinct conversations from five different domains, i.e., books, movies, soccer, music, and TV series. The questions are asked with minimal syntactic guidelines to maintain the natural factor of the questions. The questions in ConvQuestions are sourced from WikiData and the answers are provided via Web search. The questions in the dataset pose different challenges that need to be addressed including incomplete cues, anaphora, indirection, temporal reasoning, comparison, and existential.

Dataset collection Each dialog is prepared as a conversation generation task by the workers of AMT wherein they were asked to base their conversation on the five *sequential* questions from any domain of their choice. To make sure that the conversations are carried out as naturally as possible, the Turkers were asked not to interleave the questions and neither permute the order of follow-up questions to generate a large volume. Furthermore, the paraphrases of the questions were also collected to provide two versions of the questions. This would allow the data to be augmented with several interesting variations which, in turn, improves the robustness of the system. To make the dataset more closely related to real-world challenges, participants were encouraged to ask the complex questions.

Dataset analysis The dataset consists of 11,200 conversations each comprising of 5 turns. The average length of the first and follow-up questions were 9.07 and 6.20 words, respectively. Question entities and expected answers have a balanced distribution among non-human types (books, stadiums, TV series) and humans (actors, artists, authors). Context expansion is the key for finding out the correct answer in ConvQuestions as the average KG distance from the original seed to the answer is 2.30. The question type consists of characteristics such as comparisons, temporal reasoning, and anaphora, to make it more closely related to real-world challenges.

Evaluation Since each question in the dataset has exactly one or at most three correct answers, it uses standard metrics of Precision at the top rank (P@1). The other metrics include Mean Reciprocal Rank (MRR) and Hit@5. Hit@5 measures the fraction of times a correct answer is identified within the top-5 positions.

5.2 Datasets for conversational machine reading comprehension

Generally, the datasets for machine reading comprehension falls into three categories based on the type of answer they provide:

– *Multiple-choice option* datasets provide text-based multiple-choice question and expect the model to identify the right answer out of the available options. The examples of such

Table 5 A comparison of the multi-turn conversational datasets—CoQA [77] and QuAC [13] based on different characteristics as defined in their respective papers

Characteristics	CoQA	QuAC
Data source	• Passages collected from 7 diverse domains, e.g., children stories from MCTest, news articles from CNN, Wikipedia articles, etc.	• Sections from Wikipedia articles filtered in the "people" category associated with subcategories like culture, animal, geography, etc.
Conversational setup	• Questioner–answerer setting where both have access to the entire context.	• Teacher–Student setting where the teacher has access to the full context for answering, while the student has only the title and summary of the article.
Requires external knowledge?	• Yes	• No
Question type	• Factoid	• Open-ended, highly contextual
Answer type	• Free-form with an extractive rationale.	• Extractive span which can be yes/no or 'No Answer'.
Dialog acts	• No	• Yes
Max turns per dialog	15	11
Unanswerable questions	Yes	Yes
Total number of questions	126K	100K
Total number of dialogs	8K	14K

datasets include the ReAding Comprehension Dataset from Examinations (RACE) [37], Machine Comprehension Text (MCTest) [79], and MCSript [59],

– *Descriptive answer* datasets allow answers to be in any free-form text. Such datasets are useful in situations, wherein the questions are implicit and may require the use of common sense or world knowledge. The examples include MS MARCO [57] and Narrative QA [36], and
– *Span prediction* or *extractive* datasets require the model to extract the correct answer span from the given source passage. Such datasets provide better natural language understandability and easy evaluation of the task. SQuAD [75], TriviaQA [34], and NewsQA [97] are some of the popular examples of extractive datasets.

CoQA [77] and QuAC [13], the two datasets for CMRC, come under the category of span prediction datasets. Apart from these two datasets, there is another CMRC dataset, ShARC [81], which requires the understanding of a rule text to answer a few interlinked and co-referenced questions. These generated questions need to be answered using reasoning on the basis of background knowledge. However, this dataset does not really follow the definition of CMRC and is hence ignored. A summarized comparison pertaining to significant characteristics of both CoQA and QuAC is presented in Table 5.

5.2.1 CoQA

CoQA was introduced by Reddy et al. [77] to measure a machine's ability to participate in a QA style conversation. The dataset was developed with three objectives in mind. The first is the nature of questions in human conversations. In this dataset, every question except the first one is dependent on the conversation history to make it more similar to the real-life

setting of human conversation. The second goal of CoQA is to maintain the naturalness of answers in a conversation. Many existing datasets limit answers to be found in the given source passage. However, such a setting does not always ensure natural answers. In CoQA, the authors address this issue by proposing free-form answers while providing a text span from the given passage as a rationale to the answer.

The third goal of CoQA is to facilitate the development of CQA systems across multiple domains. The existing QA datasets mainly focus on a single domain which results in complications to test the generalization capabilities of the existing systems. Thus, CoQA extends its domains, i.e., each with its own data source. These domains include articles based on literature extracted from Project Gutenberg,[19] children's stories taken from MCTest [79], Wikipedia articles,[20] Reddit articles from Writing Prompt [20], middle and high school English exams taken from [37], science articles derived from Ai2 science question [103], and news articles taken from CNN [26]. Evaluation and Reddit are used for out-of-domain evaluation only.

Data collection Each conversation is prepared in a two annotator setting, i.e., one being a questioner and the other being an answerer. The platform of Amazon Mechanical Turk (AMT)[21] is used to pair workers on a passage through the ParlAI MTurk API [51] and both the annotators have full access to the passage.

Dataset analysis The dataset consists of 127K conversation turns gathered from 8K conversations over text passages. The average length of a conversation is 15 turns and each turn consists of a question and an answer. The distribution of CoQA is spread across multiple question types. Prefixes like *did*, *where*, *was*, *is*, and *does* are very frequent in the dataset. Also, almost every sector of CoQA contains co-references which shows that it is highly conversational. What makes conversations in CoQA even more humanlike is that sometimes they just feature one-word questions like "who?" or "where?" or even "why?." This shows that questions are context-dependent, and in order to answer correctly, the system needs to go through the previous history turns to understand the question.

Evaluation The main evaluation metric for the dataset is macro-average F1 score of word overlap and is computed separately for in-domain and out-of-domain as well.

5.2.2 QuAC

In an information-seeking dialog, the students keep asking their teacher questions for clarification about a particular topic. This idea forms the basis for this newly introduced dataset, Question Answering in Context (QuAC). Modeling such interrelated questions can be complex as the questions can be elliptical, highly context-dependent, and even sometimes unanswerable. To promote learning in such a challenging situation, QuAC presents a rich set of 14K crowd-sourced QA dialogs (consisting of 100K QA pairs).

Dataset collection The nature of interaction in QuAC is of student–teacher where the teacher has the access to the source paragraph. A student only provided with the heading of the paragraph aims to gain as much knowledge about its content as possible by asking multiple questions. The teacher tries to answer the questions by extracting correct answer spans from the source passage. Also, the teacher uses dialog acts as feedback to the students (i.e., may or may not ask a follow-up question) which results in more productive dialogs.

[19] https://www.gutenberg.org/.

[20] https://www.wikipedia.org/.

[21] https://www.mturk.com/.

Dataset analysis The dataset has long answers of maximum of 15 tokens which is an improvement over SQuAD and CoQA. Another factor worth noting is that frequent question types in QuAC are based on *Wh* words which makes the questions more open-ended, in contrast to the other QA datasets where questions are more factoid. Furthermore, 86% of the questions are highly contextual, i.e., they require the model to re-read the context to resolve the co-references. Out of these questions, 44% refer to entities or events in the dialog history, whereas 61% refer to the subject of the article.

Evaluation Besides evaluating the accuracy using F1 score, QuAC also utilizes human equivalence score (HEQ) to measure a system's performance by finding the percentage of exceeding or matching an average human's performance. HEQ-Q and HEQ-D are, therefore, HEQ scores with the instances as questions and dialogs, respectively.

6 Research trends and open challenges

CQA is a rapidly evolving field. This paper surveys new approaches that have been recently introduced to cater to the challenges pertaining to CQA. These CQA systems have the potential to be successfully utilized in practical applications:

- The KB-QA-based systems allow users to access a series of information via conversation without even composing complex SQL queries. From commercial perspective, these KB-QA-based systems can be employed either in open-domain QA (pertaining to worldly knowledge) or in closed-domain QA (such as in the medical field). A user does not have to access multiple sources of information, rather one agent would suffice her all information needs.
- CQA systems provide simplified conversational search (ConvSearch) setting [68] which has the strongest potential to become more popular than the traditional search engines such as Google or Bing, which unlike a user's expectations of getting a concise answer, provides a list of probable answers/solutions. These conversational systems can potentially be used for learning about a topic, planning an activity, seeking advice or guidance, and making a decision.
- The conversational agents play a significant role in facilitating smooth interaction with users. One of the conceivable applications could be customer support systems where a user does not have to go through the entire website and looks for the desired information.

As an emerging research area with many significant promising applications, CQA techniques are still not mature yet with many open issues remaining. In this section, we discuss several prominent ones:

- The role of context to be selected plays a significant role in providing accurate answers in CQA. With richer conversational scenarios, a number of contextual features need to be considered including personal context, social context, and task context. General research questions regarding contextual information in CQA include: "*What are the effective strategies and models to collect and integrate contextual information?*," "*Are knowledge graphs sufficient enough to capture and represent this information?*," and "*Do we need to incorporate the entire context or a relevant chunk would be enough to find the correct answer?*."

  Different models attempt to incorporate context in different ways. Out of all the history selection methods, the dynamic history selection mechanism proposed by Qu et al. [69] is more compelling and intuitive. As far as the flow methods are concerned, they consider

the latent representations of the entire context to deal with the varying conversational aspect. Similarly, for sequential KB-QA, the authors in [23] proposed the use of dialog manager to collect and maintain the previous utterances.

- Information-seeking behaviors need to be modeled for CQA setting as it provides users with the opportunity to obtain more information about the topics of their interests. The research questions related to information-seeking behavior that needs to be explored include: "*What optimal structure for clarification questions can be used to better understand the users' information need?*" and "*What effective strategies can be employed to design such clarification questions?*."
- Interpretability of a question plays a significant role when finding an answer for it. In the existing CQA systems, the models are anticipated to provide the answers to the questions without having to explain as to why and how they deduced an answer, making it difficult to understand the source and reason of an answer. CoQA is the only CQA dataset that provides reasoning for the provided answer. Another model, Cos-E [74], generates commonsense reasoning explanations for the deduced answer. Regardless of the fact whether or not the complete interpretability of CQA models is required, we can safely say that an understanding of the working of the internal model up to a certain extent can greatly help and improve the design of neural network systems in the future.
- Commonsense reasoning is a long-standing challenge in Conversational AI, i.e., whether it is incorporating the commonsense in dialog systems or QA systems. Commonsense reasoning refers to the ability of an individual to make day-to-day inferences by using or assuming basic knowledge about the real world. However, the CQA systems proposed so far work on pragmatic reasoning, i.e., finding the intended meaning(s) from the provided context because commonsense knowledge is often not explicitly explained in the data sources (i.e., KB-QA or CMRC dataset). Despite single-turn QA systems almost achieving human-level performance, the implementation of commonsense reasoning is still not very common. There are only a few research works that take commonsense reasoning into consideration when performing single-turn QA [30, 59]. There has been an increasing trend to incorporate commonsense reasoning into the single-turn MRC over the past few years. But when it comes to utilizing commonsense reasoning in CMRC, no successful attempt has been made. This may probably be owing to the fact that commonsense reasoning requires questions that needs some prior knowledge or background which the current CMRC datasets do not provide. When it comes to single-turn KB-QA, there are a number of prominent researches that utilize commonsense in a QA process [40, 46, 88]. Another effort was done by CoMET [6], wherein a Transformer to generate commonsense knowledge graphs was employed. Knowledge graphs like ConceptNet [91] and ATOMIC [83] have been designed to facilitate the implementation of commonsense in KB-QA systems. The field of sequential KB-QA remains untouched primarily because of the reason that the majority of existing methods lack the absence of connections between concepts [116].
- Lack of inference capability is one of the reasons why QA struggles with generating the correct answers. Most of the existing CQA systems are based on semantic relevance between question and the given context which limits a model's capability to reason. An example discussed by Liu et al. [41] depicts that provided the context, "*five people on board and two people on the ground died*," the system was not able to provide the correct answer "*seven*" to the question "*how many people died?*." Thus, how to design systems with strong inference ability is still an open issue and calls for further research.

7 Conclusion

Conversational Question Answering (CQA) systems have been emerging as a main technology to close the interactional gap between machines and humans owing to the advancements in pre-trained language modeling and the introduction of conversational datasets. This progress simplifies the development and progress of application areas such as online customer support, interactions with IoT devices in smart spaces, search engines, thus enabling CQA to realize its social and economic impacts. The effective incorporation of contextual information, the ability to infer the questions and ask efficient clarification questions are the main challenges pertaining to the field of CQA.

Our investigation of research activities over the past few years confirms the thriving expansion of this exciting field. In this survey, we have comprehensively discussed the field of CQA, which is further subdivided into (i) sequential KB-QA and (ii) Conversational Machine Reading Comprehension (CMRC). The general architecture of each of the category is decomposed into modules and prominent techniques employed in each module have been discussed. We subsequently introduced and discussed the representative datasets based on their characteristics. Finally, we discussed the potential applications of CQA and the identified future research directions that need to be explored for realizing natural conversations.

We anticipate that this literature survey will serve as a quintessence for the researchers and pave a way forward for streamlining the research in this important area.

Acknowledgements Munazza Zaib sincerely acknowledges the generous support of the Macquarie University, Sydney, Australia for funding this research work via its International Macquarie University Research Excellence Scholarship (Allocation No. 20201589). Wei Emma Zhang and Quan Z. Sheng have been partially supported by Australian Research Council (ARC) Discovery Grant DP200102298.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

1. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd international conference on learning representations, pp 01–15
2. Bao J, Duan N, Yan Z, Zhou M, Zhao T (2016) Constraint-based question answering with knowledge graph. In: Proceedings of the 26th international conference on computational linguistics, Osaka, Japan, pp 2503–2514
3. Beaver I, Freeman C, Mueen A (2020) Towards awareness of human relational strategies in virtual agents. In: Proceedings of the 34th conference on artificial intelligence, New York, pp 2602–2610
4. Bhutani N, Zheng X, Qian K, Li Y, Jagadish H (2020) Answering complex questions by combining information from curated and extracted knowledge bases. In: Proceedings of the first workshop on natural language interfaces, pp 1–10. https://doi.org/10.18653/v1/2020.nli-1.1
5. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the international conference on management of data, pp 1247–1250. https://doi.org/10.1145/1376616.1376746

6. Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y (2019) COMET: commonsense transformers for automatic knowledge graph construction. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy, pp 4762–4779. https://doi.org/10.18653/v1/P19-1470
7. Bouziane A, Bouchiha D, Doumi N, Malki M (2015) Question answering systems: survey and trends. Procedia Comput Sci 73:366–375. https://doi.org/10.1016/j.procs.2015.12.005
8. Budzianowski P, Wen TH, Tseng BH, Casanueva I, Stefan U, Osman R, Gašić M (2018) MultiWOZ: a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: Proceedings of the conference on empirical methods in natural language processing, Brussels, Belgium, pp 5016–5026. https://doi.org/10.18653/v1/D18-1547
9. Cascante-Bonilla P, Sitaraman K, Luo M, Ordonez V (2019) Moviescope: large-scale analysis of movies using multiple modalities. arXiv:1908.03180
10. Chen D, Fisch A, Weston J, Bordes A (2017) Reading Wikipedia to answer open-domain questions. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Vancouver, Canada, pp 1870–1879. https://doi.org/10.18653/v1/P17-1171
11. Chen Y, Wu L, Zaki MJ (2020) Graphflow: exploiting conversation flow with graph neural networks for conversational machine comprehension. In: Proceedings of the 29th international joint conference on artificial intelligence, pp 1230–1236
12. Cheng J, Reddy S, Saraswat V, Lapata M (2019) Learning an executable neural semantic parser. Computat Linguist. https://doi.org/10.1162/coli_a_00342
13. Choi E, He H, Iyyer M, Yatskar M, Yih W, Choi Y, Liang P, Zettlemoyer L (2018) QuAC: question answering in context. In: Proceedings of the conference on empirical methods in natural language processing, Brussels, Belgium, pp 2174–2184. https://doi.org/10.18653/v1/D18-1241
14. Christmann P, Roy RS, Abujabal A, Singh J, Weikum G (2019) Look before you hop: conversational question answering over knowledge graphs using judicious context expansion. In: Proceedings of the 28th ACM international conference on information and knowledge management, Beijing, China, pp 729–738. https://doi.org/10.1145/3357384.3358016
15. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. In: Proceedings of the 37th international conference on neural information processing systems, Montreal, Canada, pp 01–09
16. Cui L, Huang S, Wei F, Tan C, Duan C, Zhou M (2017a) Superagent: a customer service chatbot for e-commerce websites. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Vancouver, Canada, pp 97–102. https://doi.org/10.18653/v1/P17-4017
17. Cui W, Xiao Y, Wang H, Song Y, Hwang S, Wang W (2017b) KBQA: learning question answering over QA corpora and knowledge bases. Proc VLDB Endow 10(5), 565–576. https://doi.org/10.14778/3055540.3055549 https://doi.org/10.14778/3055540.3055549 https://doi.org/10.14778/3055540.3055549
18. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies, Minneapolis, USA, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
19. Dhingra B, Li L, Li X, Gao J, Chen Y, Ahmed F, Deng L (2017) Towards end-to-end reinforcement learning of dialogue agents for information access. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Vancouver, Canada, pp 484–495. https://doi.org/10.18653/v1/P17-1045
20. Fan A, Lewis M, Dauphin YN (2018) Hierarchical neural story generation. In: Proceedings of the 56th annual meeting of the association for computational linguistics, Melbourne, Australia, pp 889–898. https://doi.org/10.18653/v1/P18-1082
21. Fu B, Qiu Y, Tang C, Li Y, Yu H, Sun J (2020) A survey on complex question answering over knowledge base: recent advances and challenges. arXiv:2007.13069
22. Gao J, Galley M, Li L (2019) Neural approaches to conversational AI. Found Trends Inf Retr 13(2–3):127–298. https://doi.org/10.1561/1500000074
23. Guo D, Tang D, Duan N, Zhou M, Yin J (2018) Dialog-to-action: conversational question answering over a large-scale knowledge base. In: Proceedings of the 32nd international conference on neural information processing systems, Montréal, Canada, pp 2946–2955
24. Gupta S, Rawat BPS, Yu H (2020) Conversational machine comprehension: a literature review. In: Proceedings of the 28th international conference on computational linguistics, pp 2739–2753
25. Gur I, Hewlett D, Lacoste A, Jones L (2017) Accurate supervised and semi-supervised machine reading for long documents. In: Proceedings of the conference on empirical methods in natural language processing, Copenhagen, Denmark, pp 2011–2020. https://doi.org/10.18653/v1/D17-1214

26. Hermann KM, Kociský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. In: Proceedings of the 28th international conference on neural information processing systems, Montréal, Canada, pp 1693–1701
27. Higashinaka R, Isozaki H (2008) Corpus-based question answering for why-questions. In: Proceedings of the 3rd international joint conference on natural language processing, Hyderabad, India, pp 01–08
28. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
29. Huang H, Choi E, Yih W (2019a) FlowQA: grasping flow in history for conversational machine comprehension. In: Proceedings of the 7th international conference on learning representations, New Orleans, LA, USA, pp 01–08
30. Huang L, Bras RL, Bhagavatula C, Choi Y (2019b) Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In: Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Hong Kong, China, pp 2391–2401
31. Iyyer M, Boyd-Graber JL, Claudino LMB, Socher R, III HD (2014) A neural network for factoid question answering over paragraphs. In: Proceedings of the conference on empirical methods in natural language processing, pp 633–644. https://doi.org/10.3115/v1/D14-1070
32. Iyyer M, Yih W, Chang M (2017) Search-based neural structured learning for sequential question answering. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Vancouver, Canada, pp 1821–1831. https://doi.org/10.18653/v1/P17-1167
33. Jiang K, Wu D, Jiang H (2019) FreebaseQA: a new factoid QA data set matching trivia-style question-answer pairs with Freebase. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies, Napa Valley, California, USA, pp 318–323. https://doi.org/10.1145/1376616.1376746
34. Joshi M, Choi E, Weld DS, Zettlemoyer L (2017) TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Minneapolis, Minnesota, pp 1601–1611. https://doi.org/10.18653/v1/N19-1028
35. Kacupaj E, Plepi J, Singh K, Thakkar H, Lehmann J, Maleshkova M (2021) Conversational question answering over knowledge graphs with transformer and graph attention networks. In: Proceedings of the 16th conference of the european chapter of the association for computational linguistics, pp 850–862. https://doi.org/10.18653/v1/2021.eacl-main.72
36. Kociský T, Schwarz J, Blunsom P, Dyer C, Hermann KM, Melis G, Grefenstette E (2018) The narrative QA reading comprehension challenge. Trans Assoc Comput Linguist. https://doi.org/10.1162/tacl_a_00023
37. Lai G, Xie Q, Liu H, Yang Y, Hovy EH (2017) RACE: large-scale reading comprehension dataset from examinations. In: Proceedings of the conference on empirical methods in natural language processing, Copenhagen, Denmark, pp 785–794. https://doi.org/10.18653/v1/D17-1082
38. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) ALBERT: a lite BERT for self-supervised learning of language representations. In: Proceedings of the 8th international conference on learning representations
39. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, Van Kleef P, Auer S et al (2015) DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2):167–195
40. Lin BY, Chen X, Chen J, Ren X (2019) KagNet: knowledge-aware graph networks for commonsense reasoning. In: Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Hong Kong, China, pp 2829–2839. https://doi.org/10.18653/v1/D19-1282
41. Liu S, Zhang S, Zhang X, Wang H (2019) R-Trans: RNN transformer network for Chinese machine reading comprehension. IEEE Access 7:27736–27745. https://doi.org/10.1109/ACCESS.2019.2901547
42. Liu S, Zhang X, Zhang S, Wang H, Zhang W (2019) Neural machine reading comprehension: methods and trends. Appl Sci 9(18):3698. https://doi.org/10.3390/app9183698
43. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019b) RoBERTa: a robustly optimized Bert pretraining approach. arXiv:1907.11692
44. Lu X, Pramanik S, Roy RS, Abujabal A, Wang Y, Weikum G (2019) Answering complex questions by joining multi-document evidence with quasi knowledge graphs. In: Proceedings of the 42nd international conference on research and development in information retrieval, Paris, France, pp 105–114. https://doi.org/10.1145/3331184.3331252

45. Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the conference on empirical methods in natural language processing, Lisbon, Portugal, pp 1412–1421. https://doi.org/10.18653/v1/D15-1166
46. Lv S, Guo D, Xu J, Tang D, Duan N, Gong M, Shou L, Jiang D, Cao G, Hu S (2020) Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In: Proceeding of the 34th conference on artificial intelligence, New York, USA, pp 8449–8456
47. Martinez-Gil J (2015) Automated knowledge base management: a survey. Comput Sci Rev 18:1–9. https://doi.org/10.1016/j.cosrev.2015.09.001
48. McCann B, Bradbury J, Xiong C, Socher R (2017) Learned in translation: contextualized word vectors. In: Proceedings of the 31st international conference on neural information processing systems, Long Beach, California, USA, pp 6294–6305
49. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Proceedings of the 1st international conference on learning representations, Scottsdale, Arizona, pp 01–12
50. Miller A, Fisch A, Dodge J, Karimi AH, Bordes A, Weston J (2016) Key-value memory networks for directly reading documents. In: Proceedings of the conference on empirical methods in natural language processing, Austin, Texas, pp 1400–1409. https://doi.org/10.18653/v1/D16-1147
51. Miller AH, Feng W, Batra D, Bordes A, Fisch A, Lu J, Parikh D, Weston J (2017) ParlAI: a dialog research software platform. In: Proceedings of the conference on empirical methods in natural language processing, Copenhagen, Denmark, pp 79–84. https://doi.org/10.18653/v1/D17-2014
52. Mishra A, Jain SK (2016) A survey on question answering systems with classification. J King Saud Univ Comput Inf Sci 28(3):345–361. https://doi.org/10.1016/j.jksuci.2014.10.007
53. Mitchell T, Cohen W, Hruschka E, Talukdar P, Yang B, Betteridge J, Carlson A, Dalvi B, Gardner M, Kisiel B et al (2018) Never-ending learning. Commun ACM 61(5):103–115. https://doi.org/10.1145/3191513
54. Monz C (2011) Machine learning for query formulation in question answering. Nat Lang Eng 17(4):425–454. https://doi.org/10.1017/S1351324910000276
55. Müller T, Piccinno F, Shaw P, Nicosia M, Altun Y (2019) Answering conversational questions on structured data without logical forms. In: Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Hong Kong, China, pp 5901–5909
56. Nallapati R, Zhou B, dos Santos CN, Gülçehre Ç, Xiang B (2016) Abstractive text summarization using sequence-to-sequence rnns and beyond. In: Proceedings of the 20th SIGNLL conference on computational natural language learning, Berlin, Germany, pp 280–290. https://doi.org/10.18653/v1/K16-1028
57. Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, Deng L (2016) MS MARCO: a human generated machine reading comprehension dataset. In: Proceedings of the 30th annual conference on neural information processing systems, Barcelona, Spain, pp 01–11
58. Ohsugi Y, Saito I, Nishida K, Asano H, Tomita J (2019) A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy, pp 11–17. https://doi.org/10.18653/v1/W19-4102
59. Ostermann S, Modi A, Roth M, Thater S, Pinkal M (2018) MCScript: a novel dataset for assessing machine comprehension using script knowledge. In: Proceedings of the 11th international conference on language resources and evaluation, Miyazaki, Japan, pp 01–08
60. Pasupat P, Liang P (2015) Compositional semantic parsing on semi-structured tables. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, Beijing, China, pp 1470–1480. https://doi.org/10.3115/v1/P15-1142
61. Peng B, Zhu C, Li C, Li X, Li J, Zeng M, Gao J (2020) Few-shot natural language generation for task-oriented dialog. In: Proceedings of the conference on empirical methods in natural language processing, pp 172–182
62. Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation. In: Proceedings of the conference on empirical methods in natural language processing, Doha, Qatar, pp 1532–1543. https://doi.org/10.3115/v1/D14-1162
63. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the conference of the north American chapter of the association for computational linguistics: human language technologies, New Orleans, Louisiana, pp 2227–2237. https://doi.org/10.18653/v1/N18-1202

64. Pinto D, Branstein M, Coleman R, Croft WB, King M, Li W, Wei X (2002) Quasm: a system for question answering using semi-structured data. In: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, Oregon, USA, pp 46–55. https://doi.org/10.1145/544220.544228
65. Plepi J, Kacupaj E, Singh K, Thakkar H, Lehmann J (2021) Context transformer with stacked pointer networks for conversational question answering over knowledge graphs. In: Proceedings of the 18th international semantic web conference, Springer, vol 12731, pp 356–371
66. Qi P, Lin X, Mehr L, Wang Z, Manning CD (2019) Answering complex open-domain questions through iterative query generation. In: Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Florence, Italy, pp 2590–2602
67. Qiu M, Huang X, Chen C, Ji F, Qu C, Wei W, Huang J, Zhang Y (2021) Reinforced history backtracking for conversational question answering. In: Proceedings of the 35th conference on artificial intelligence
68. Qu C, Yang L, Qiu M, Croft WB, Zhang Y, Iyyer M (2019a) BERT with history answer embedding for conversational question answering. In: Proceedings of the 42nd international conference on research and development in information retrieval, Paris France, pp 1133–1136. https://doi.org/10.1145/3331184.3331341
69. Qu C, Yang L, Qiu M, Zhang Y, Chen C, Croft WB, Iyyer M (2019b) Attentive history selection for conversational question answering. In: Proceedings of the 28th international conference on information and knowledge management, Beijing, China, pp 1391–1400
70. Qu C, Yang L, Chen C, Qiu M, Croft WB, Iyyer M (2020) Open-retrieval conversational question answering. In: Proceedings of the 43rd international conference on research and development in information retrieval, pp 539–548. https://doi.org/10.1145/3397271.3401110
71. Qu C, Yang L, Chen C, Croft WB, Krishna K, Iyyer M (2021) Weakly-supervised open-retrieval conversational question answering. In: Proceedings of the 43rd European conference on IR research, pp 529–543
72. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training
73. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
74. Rajani NF, McCann B, Xiong C, Socher R (2019) Explain yourself! Leveraging language models for commonsense reasoning. In: Proceedings of the 57th conference of the association for computational linguistics, Florence, Italy, pp 4932–4942. https://doi.org/10.18653/v1/P19-1487
75. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100, 000+ questions for machine comprehension of text. In: Proceedings of the conference on empirical methods in natural language processing, Austin, Texas, pp 2383–2392. https://doi.org/10.18653/v1/D16-1264
76. Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: unanswerable questions for SQuAD. In: Proceedings of the 56th annual meeting of the association for computational linguistics, Melbourne, Australia, pp 784–789. https://doi.org/10.18653/v1/P18-2124
77. Reddy S, Chen D, Manning CD (2019) CoQA: a conversational question answering challenge. Trans Assoc Comput Linguist 7:249–266. https://doi.org/10.1162/tacl_a_00266
78. Ren L, Xie K, Chen L, Yu K (2018) Towards universal dialogue state tracking. In: Proceedings of the conference on empirical methods in natural language processing, Brussels, Belgium, pp 2780–2786. https://doi.org/10.18653/v1/D18-1299
79. Richardson M, Burges CJC, Renshaw E (2013) MCTest: a challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the conference on empirical methods in natural language processing, Seattle, Washington, USA, pp 193–203
80. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536. https://doi.org/10.1038/323533a0
81. Saeidi M, Bartolo M, Lewis PSH, Singh S, Rocktäschel T, Sheldon M, Bouchard G, Riedel S (2018) Interpretation of natural language rules in conversational machine reading. In: Proceedings of the conference on empirical methods in natural language processing, Brussels, Belgium, pp 2087–2097. https://doi.org/10.18653/v1/D18-1233
82. Saha A, Pahuja V, Khapra MM, Sankaranarayanan K, Chandar S (2018) Complex sequential question answering: towards learning to converse over linked question answer pairs with a knowledge graph. In: Proceedings of the 32nd conference on artificial intelligence, New Orleans, Louisiana, USA, pp 705–713
83. Sap M, Le Bras R, Allaway E, Bhagavatula C, Lourie N, Rashkin H, Roof B, Smith NA, Choi Y (2019) ATOMIC: an atlas of machine commonsense for if-then reasoning. In: Proceedings of the 33rd conference on artificial intelligence, Hawaii, USA, vol 33, pp 3027–3035. https://doi.org/10.1609/aaai.v33i01.33013027

84. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2008) The graph neural network model. IEEE Trans Neural Netw 20(1):61–80
85. See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Vancouver, Canada, pp 1073–1083
86. Seo MJ, Kembhavi A, Farhadi A, Hajishirzi H (2017) Bidirectional attention flow for machine comprehension. In: 5th International conference on learning representations, ICLR 2017. Toulon, France, pp 01–13
87. Serban I, Sordoni A, Bengio Y, Courville A, Pineau J (2016) Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the 30th conference on artificial intelligence, Phoenix, Arizona, USA, pp 3776–3784
88. Sharma A, Goolsbey KM (2019) Simulation-based approach to efficient commonsense reasoning in very large knowledge bases. In: Proceedings of the 33rd conference on artificial intelligence, Hawaii, USA, pp 1360–1367. https://doi.org/10.1609/aaai.v33i01.33011360
89. Shen D, Klakow D (2006) Exploring correlation of dependency relation paths for answer extraction. In: Proceedings of the 44th annual meeting of the association for computational linguistics, ACL 2006, Sydney, Australia, pp 889-896. https://doi.org/10.3115/1220175.1220287
90. Shen T, Geng X, Qin T, Guo D, Tang D, Duan N, Long G, Jiang D (2019) Multi-task learning for conversational question answering over a large-scale knowledge base. In: Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Hong Kong, China, pp 2442–2451. https://doi.org/10.18653/v1/D19-1248
91. Speer R, Chin J, Havasi C (2017) ConceptNet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the 31st conference on artificial intelligence, San Francisco, California, USA, pp 4444–4451
92. Suhr A, Iyer S, Artzi Y (2018) Learning to map context-dependent sentences to executable formal queries. In: Proceedings of the conference of the North American Chapter of the association for computational linguistics: human language technologies, New Orleans, Louisiana, pp 2238–2249. https://doi.org/10.18653/v1/N18-1203
93. Sun R, Cao X, Zhao Y, Wan J, Zhou K, Zhang F, Wang Z, Zheng K (2020) Multi-modal knowledge graphs for recommender systems. In: Proceedings of the 29th ACM international conference on information and knowledge management, pp 1405–1414. https://doi.org/10.1145/3340531.3411947
94. Suster S, Daelemans W (2018) CliCR: a dataset of clinical case reports for machine reading comprehension. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies, New Orleans, Louisiana, pp 1551–1563. https://doi.org/10.18653/v1/N18-1140
95. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst Montreal Canada 27:3104–3112
96. Tian Z, Yan R, Mou L, Song Y, Feng Y, Zhao D (2017) How to make context more useful? an empirical study on context-aware neural conversational models. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Vancouver, Canada, pp 231–236. https://doi.org/10.18653/v1/P17-2036
97. Trischler A, Wang T, Yuan X, Harris J, Sordoni A, Bachman P, Suleman K (2017) NewsQA: a machine comprehension dataset. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Vancouver, Canada, pp 191–200. https://doi.org/10.18653/v1/W17-2623
98. Trivedi P, Maheshwari G, Dubey M, Lehmann J (2017) LC-QuaD: a corpus for complex question answering over knowledge graphs. In: Proceedings of the 16th international semantic web conference, pp 210–218
99. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems, vol 30, Long Beach, California, USA, pp 5998-6008,
100. Velickovic P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph attention networks. In: Proceedings of the 6th international conference on learning representations, Vancouver, Canada
101. Vinyals O, Fortunato M, Jaitly N (2015) Pointer networks. In: Proceedings of the 29th international conference on neural information processing systems, vol 28, Montréal, Canada, pp 2692–2700
102. Wang H, Zhang X, Ma S, Sun X, Wang H, Wang M (2018) A neural question answering model based on semi-structured tables. In: Proceedings of the 27th international conference on computational linguistics, Santa Fe, New Mexico, USA, pp 1941–1951
103. Welbl J, Liu NF, Gardner M (2017) Crowdsourcing multiple choice science questions. In: Proceedings of the conference on empirical methods in natural language processing, Copenhagen, Denmark, pp 94–106. https://doi.org/10.18653/v1/W17-4413

104. Wen TH, Vandyke D, Mrkšić N, Gašić M, Rojas-Barahona LM, Su PH, Ultes S, Young S (2017) A network-based end-to-end trainable task-oriented dialogue system. In: Proceedings of the 15th conference of the european chapter of the association for computational linguistics, Valencia, Spain, pp 438–449. https://doi.org/10.18653/v1/E17-1042
105. Wu J, Li M, Lee CH (2015) A probabilistic framework for representing dialog systems and entropy-based dialog management through dynamic stochastic state evolution. IEEE/ACM Trans Audio Speech Lang Process 23(11):2026–2035. https://doi.org/10.1109/TASLP.2015.2462712
106. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144
107. Xiong W, Li X, Iyer S, Du J, Lewis P, Wang WY, Mehdad Y, Yih S, Riedel S, Kiela D, Oguz B (2021) Answering complex open-domain questions with multi-hop dense retrieval. In: Proceedings of the 9th international conference on learning representations, pp 01 –19
108. Yang Y, Yih Wt, Meek C (2015) WikiQA: a challenge dataset for open-domain question answering. In: Proceedings of the conference on empirical methods in natural language processing, Lisbon, Portugal, pp 2013–2018. https://doi.org/10.18653/v1/D15-1237
109. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. In: Proceedings of the 33rd international conference on neural information processing systems, vol 32, Vancouver, Canada, pp 5754–5764
110. Yatskar M (2019) A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. In: Proceedings of the conference of the North American Chapter of the association for computational linguistics: human language technologies, Minneapolis, Minnesota, pp 2318–2323. https://doi.org/10.18653/v1/N19-1241
111. Yeh Y, Chen Y (2019) FlowDelta: modeling flow information gain in reasoning for conversational machine comprehension. In: Proceedings of the conference on empirical methods in natural language processing, Hong Kong, China, pp 86–90. https://doi.org/10.18653/v1/D19-5812
112. Zaib M, Sheng QZ, Zhang WE (2020) A short survey of pre-trained language models for conversational AI: A new age in NLP. In: Proceedings of the Australasian computer science week multiconference 2020, Melbourne, Australia, pp 1–4. https://doi.org/10.1145/3373017.3373028
113. Zaib M, Tran DH, Sagar S, Mahmood A, Zhang WE, Sheng QZ (2021) BERT-CoQAC: BERT-based conversational question answering in context. In: Parallel architectures, algorithms and programming, pp 47–57. https://doi.org/10.1007/978-981-16-0010-4_5
114. Zellers R, Bisk Y, Schwartz R, Choi Y (2018) SWAG: a large-scale adversarial dataset for grounded commonsense inference. In: Proceedings of the conference on empirical methods in natural language processing, Brussels, Belgium, pp 93–104
115. Zhang Y, Chen X, Ai Q, Yang L, Croft WB (2018) Towards conversational search and recommendation: system ask, user respond. In: Proceedings of the 27th ACM international conference on information and knowledge management, Torino, Italy, pp 177–186. https://doi.org/10.1145/3269206.3271776
116. Zhong W, Tang D, Duan N, Zhou M, Wang J, Yin J (2019) Improving question answering by commonsense-based pre-training. In: Proceedings of the 8th international natural language processing and chinese computing conference, Dunhuang, China, pp 16–28. https://doi.org/10.1007/978-3-030-32233-5_2
117. Zhu C, Zeng M, Huang X (2018) SDNet: contextualized attention-based deep network for conversational question answering. arXiv:1812.03593
118. Zhu C, Zeng M, Huang X (2019) Multi-task learning for natural language generation in task-oriented dialogue. In: Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Hong Kong, China, pp 1261–1266. https://doi.org/10.18653/v1/D19-1123
119. Zou X (2020) A survey on application of knowledge graph. J Phys Conf Ser 1487:012016. https://doi.org/10.1088/1742-6596/1487/1/012016

Munazza Zaib received the B.E. degree in Software Engineering from Mehran University of Engineering Technology, Jamshoro, Pakistan and a Master's degree in Computing from Macquarie University, Sydney, Australia in 2013 and 2020, respectively. She is currently pursuing Ph.D. degree at the School of Computing, Macquarie University, Sydney, Australia. Before moving to Macquarie University, Munazza, worked as a faculty member at the Department of Software Engineering, Mehran University of Engineering Technology, Khairpur Mirs, Pakistan, from 2016 to 2019. Her current research interests include natural language processing, information retrieval, question answering systems, and their applications.

Wei Emma Zhang currently works as a Lecturer at the School of Computer Science, the University of Adelaide. She is also an Honorary lecturer at School of Computing, Macquarie University. She got her PhD in Computer Science in 2017 from The University of Adelaide. Her research interests include text mining, natural language processing, deep learning, information retrieval, and Internet of Things. She has 90+ publications as edited books and proceedings, refereed book chapters, and refereed technical papers in journals and conferences including ACM CSUR, IEEE COMST, ACM TIST, ACM TOIT, ACM TOSN, WWWJ, TBD, CACM, ACL, ACM SIGIR, WWW, EDBT, CIKM, ECCV, ICSOC, and CAiSE. Her PhD thesis had been published by Springer as a monograph. She is also active in academic community services for more than 30 conferences and journals in artificial intelligence, data mining, information retrieval, and Internet of Things. She is the member of ACM and IEEE.

Quan Z. Sheng is a full Professor and Head of School of Computing at Macquarie University. Before moving to Macquarie, he spent 10 years at School of Computer Science, the University of Adelaide, serving in a number of senior leadership roles including acting Head and Deputy Head of School of Computer Science. He holds a PhD degree in computer science from the University of New South Wales and did his post doc as a research scientist at CSIRO ICT Centre. His research interests include the Internet of Things, big data analytics, knowledge discovery, and Internet technologies. Professor Sheng is ranked by Microsoft Academic as one of the Most Impactful Authors in Services Computing (ranked Top 5 All Time). He is the recipient of the AMiner Most Influential Scholar Award on IoT (2007–2017), ARC Future Fellowship (2014), Chris Wallace Award for Outstanding Research Contribution (2012), and Microsoft Research Fellowship (2003).

Adnan Mahmood holds a PhD in Computer Science and is currently a Postdoctoral Research Fellow at the School of Computing, Macquarie University, Australia. Before moving to Macquarie University, Adnan has spent a considerable number of years in the academic and research settings of Republic of Ireland, South Korea, Malaysia, Pakistan, and People's Republic of China. His research interests include software-defined networks, intelligent transportation systems, Internet of Things (primarily Internet of Vehicles), trust management, and next-generation heterogeneous wireless networks. Adnan, besides, serve on the Technical Program Committees of a number of reputed International Conferences. He is a member of the IEEE, IET, and the ACM

Yang Zhang is currently a Cotutelle PhD Student at (a) the School of Information Science, Wuhan University, Wuhan, P.R.China and (b) the School of Computing, Macquarie University, Sydney, Australia. His research interests include, but are not limited to, natural language processing, citation analysis, knowledge discovery, and informetrics. Yang regularly publishes in various International Conferences and Journals of repute and has also contributed to projects of national significance funded by the National Social Science Foundation of People's Republic of China.