A Survey of Document Grounded Dialogue Systems (DGDS)

LONGXUAN MA, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China

WEI-NAN ZHANG, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China

MINGDA LI, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China

TING LIU, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China

Dialogue system (DS) attracts great attention from industry and academia because of its wide application prospects. Researchers usually divide the DS according to the function. However, many conversations require the DS to switch between different functions. For example, movie discussion can change from chit-chat to QA, the conversational recommendation can transform from chit-chat to recommendation, etc. Therefore, classification according to functions may not be enough to help us appreciate the current development trend. We classify the DS based on background knowledge. Specifically, study the latest DS based on the unstructured document(s). We define Document Grounded Dialogue System (DGDS) as **the DS that the dialogues are centering on the given document(s)**. The DGDS can be used in scenarios such as talking over merchandise against product Manual, commenting on news reports, etc. We believe that extracting unstructured document(s) information is the future trend of the DS because a great amount of human knowledge lies in these document(s). The research of the DGDS not only possesses a broad application prospect but also facilitates AI to better understand human knowledge and natural language. We analyze the classification, architecture, datasets, models, and future development trends of the DGDS, hoping to help researchers in this field.

CCS Concepts: • **General and reference → Surveys and overviews**.

Additional Key Words and Phrases: Dialogue System, Document Grounded, Chit-Chat, Conversational Reading Comprehension

ACM Reference Format:

Longxuan Ma, Wei-Nan Zhang, Mingda Li, and Ting Liu. 2019. A Survey of Document Grounded Dialogue Systems (DGDS). 1, 1 (April 2019), 30 pages.

1 INTRODUCTION

For a long time, researchers have been devoting themselves to develop a Dialogue System (DS) that can communicate with human beings naturally. Early DS such as Eliza [110], Parry [15], and Alice [23] attempted to simulate human

Authors' addresses: Longxuan Ma, lxma@ir.hit.edu.cn, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, 92, Xidazhi Road, Nangang Qu, Harbin, Heilongjiang, China; Wei-Nan Zhang, wnzhang@ir.hit.edu.cn, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, 92, Xidazhi Road, Nangang Qu, Harbin, Heilongjiang, China; Mingda Li, mdli@ir.hit.edu.cn, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, 92, Xidazhi Road, Nangang Qu, Harbin, Heilongjiang, China; Ting Liu, tliu@ir.hit.edu.cn, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, 92, Xidazhi Road, Nangang Qu, Harbin, Heilongjiang, China.

Manuscript submitted to ACM

behaviors in conversations, and challenged various forms of the Turing Test [104]. They worked well but only in constrained environments, an open-domain DS remained an elusive task until recently [37]. Then works focused on the task-oriented DS such as DARPA [107–109] arose, they performed well only within domains that have well-defined schemas.

Although task-oriented DS and open-domain DS are originally developed for different purposes, Gao et al. [26] regarded both of them can be designed as an optimal decision-making process, whose goal is to maximize the expected reward. The reward of the former is easier to define and optimize than that of the latter. In past years, some researchers [1, 20, 121, 134] have begun to explore technologies to integrate them. Dodge et al. [20] investigated 5 different tasks (QA, Dialogue, Recommendation, etc.) with a Memory Network [114]. Akasaki and Kaji [1] proposed a dataset to distinguish whether the user is having a chat or giving a request to the chatting machine. Yan et al. [121] presented a general solution to the task-oriented DS for online shopping. The goal is to help users complete various purchase-related tasks, such as searching products and answering questions, just like the dialogue between normal people. Zhao et al. [134] proposed a task-oriented dialogue agent based on the encoder-decoder model with chatting capability. Subsequently, Ghazvininejad et al. [27] presented a fully data-driven and knowledge-grounded neural conversation model aimed at producing more contentful responses without slot filling. These works represented steps that build end-to-end DS in scenarios beyond a single function.

In the past few years, a great effort has been made to develop virtual assistants such as AppleâĂŹs Siri[1], MicrosoftâĂŹs Cortana[2], AmazonâĂŹs Alexa[3] and Google Assistant[4]. These applications are capable of answering a wide range of questions on mobile devices. In addition to passively responding to user requests, they also proactively predict users' demands and provide in-time assistance such as reminding of an upcoming event or recommending a useful service without receiving explicit commands [87]. Meanwhile, social bots designed to meet the users' emotional needs showed great development potential. Since its launch in 2014, Microsoft's XiaoIce System[5] has attracted millions of people on various topics for long time interlocution [93, 139]. Inspired by the Turing test, it is designed to test the ability of social bots to provide coherent, relevant, interesting and interactive communication and to keep user's participation within the possible range. In 2016, the Alexa Prize challenge was proposed to advance the research and development of social bots that are able to converse coherently and engagingly with humans on popular topics such as sports, politics, and entertainment, for at least 20 minutes [80]. The virtual assistants and social bots usually consist a natural hierarchy: a top-level process selecting what the agent is about to activate for a particular subtask (e.g., answering a question, scheduling a meeting, providing a recommendation or just having a casual chat), and a low-level process choosing primitive actions to complete the subtask. However, due to the difficulty of natural language understanding (NLU) and natural language generation (NLG), the universal intelligence embodied in these systems still lags behind human beings.

As we introduced, the modern DS pays more attention to the integration of multiple functions to improve the interactive experience, which makes the function-based classification of the DS insufficient to reflect the current progress. In this paper, we divide the DS by background knowledge which is defined as unstructured text knowledge besides the content of the conversation and can be used during the dialogue. Specifically, we focus on the DS based on the unstructured document(s), namely Document Grounded Dialogue System (DGDS). The DGDS tries to establish a

[1]https://www.apple.com/ios/siri/
[2]https://www.microsoft.com/en-us/cortana/
[3]https://developer.amazon.com/alexa/
[4]https://assistant.google.com/
[5]https://www.msXiaoIce.com/

conversation mode in which relevant information can be obtained from the given document(s). Despite the increasing efforts with introducing external structured or unstructured knowledge into the conversation to generate more informative replies in the DS, the DGDS is different from works [27, 115, 122] that first retrieve a set of candidate facts or responses and then generate the response by treating the retrieval results as additional knowledge, because the background information the DGDS used is the document(s) and the internal structure of the text need to be considered. We also do not take into account the process of initial selection of candidate documents from a larger knowledge environment (e.g. The Web), because it usually adopts common IR technology (e.g. Keywords, TF-IDF) to find some candidate document(s) [6, 18, 56, 98, 99], and assumes that ideally, the top-N results cover the most related knowledge. The dialogue history of the DGDS must be based on the same pre-determined document(s), choosing new document(s) as knowledge sources according to the dialogue history should not be counted.

Besides chit-chat with external knowledge, the DGDS also shares some common features with the MRC. In fact, if a single round QA is counted as a conversation, the DGDS could cover some popular MRC[6] tasks [34, 35, 44, 67, 79, 102]. Some DGDS models [66, 73, 83] also employ MRC models as baselines [89] or component [53]. However, single round QA does not need to pay attention to the history of dialogue, resulting in obvious differences with the multi-turn DGDS in the modeling process, memory ability, reasoning ability, evaluation methods, etc. We only include multi-turn Conversational Reading Comprehension (CRC) in the DGDS because it is consistent with our definition. Besides, the single-turn MRC review research has been systematic and comprehensive [52, 74, 130**?**], but the survey on the multi-turn CRC is scarce.

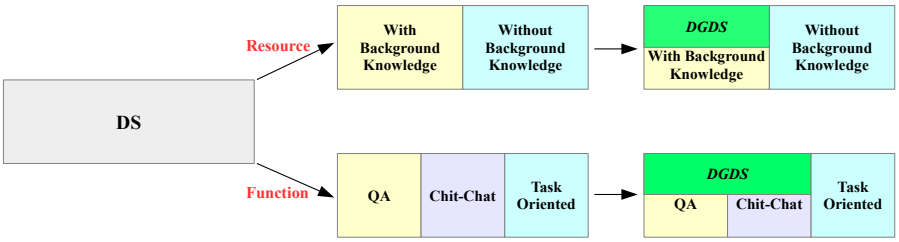

Fig. 1. The position of the DGDS in Dialogue System.

In this paper, we studied the multi-turn DGDS, including Conversational Reading Comprehension (CRC) where the conversations are QA mode and Document-Based Dialogue (DBD) where the dialogues are chit-chat form. The scope of the DGDS in DS is demonstrated in Figure 1. Since most of the DGDS datasets are released from 2018 to 2019, this paper mainly focuses on the relevant deep learning models. As far as we know, we are the first to make a systematic review of this field. We believe that incorporating unstructured document(s) information into response generation is the inevitable trend of the open domain DS because a large number of human knowledge is hidden in these document(s). The research of the DGDS can better assistant AI in using human knowledge and improving AI's understanding of natural language.

The structure of the paper is as follows:

- In Chapter **Introduction**, we give a brief history introduction of the DS and the DGDS.

[6]in the broad sense, Cloze Tests, Multiple Choice, Span Extraction, Free Answering, KBQA, CQA, etc., are all belong to the MRC task, while in this article we refer to the MRC as the QA tasks based on document(s).

- In Chapter **Comparison**, we analyze the difference between the DGDS and three classifications (task-oriented, chit-chat, and QA) of the DS from different perspectives then compare the CRC with the DBD.
- In Chapter **Architecture**, the main architecture of the DGDS models are outlined.
- In Chapter **Datasets**, we review the DGDS datasets that have been released.
- In Chapter **CRC Models** and **DBD Models**, we discuss the CRC and the DBD approaches based on the pre-defined architecture respectively.
- In Chapter **Future Work**, we put forward the promising research directions in both fundamental and technical aspects.

2 COMPARISON

In this chapter, we demonstrate the distinctions among the DGDS and the traditional DS categories and analyze the differences between the CRC and the DBD in detail. We first define some concepts frequently used in this research field in Table 1 to avoid confusion.

Table 1. Related concepts in DGDS.

Concept	Definition
Document(s)	The text used for discussion in a conversation, the content of which has an inseparable coherent logical relationship.
QA	One interlocutor asks questions, the other one provides answers if exist.
Turn	In QA, one turn is one question or one answer, in chit-chat, one turn is consecutive utterances from one speaker.
Exchange	In QA, each exchange is one QA pair. In chit-chat, any two consecutive turns make one exchange.
Dialogue/ Conversation	Multiple turns/exchanges constitute a dialogue/conversation.
Utterance	A sequence of sentences from one speaker in one turn.
Context	All the utterances but the current one. Context usually means the conversation history.
Evidence	Text segment(s) containing the information for response generation.
Agent/Bot	Model trained as dialogue partner.
User	Real human participating in the dialogue.

2.1 Differences Between the DGDS and the other DS

Accompanied by the rapid development, surveys of the DS nowadays attempt to analyze from different perspectives. In view of task or non-task based DS, Chen et al. [7] analyzed the retrieval, generation and hybridization models. Sarikaya [87] summarized the technology behind personal digital assistants, such as the system architecture and key components. Serban et al. [90] classified datasets for building DS from a data-driven perspective. Shum et al. [93] outlined the advantages and disadvantages of the current social chatbots. Yan [120] summarized the non-task-oriented chit-chat bots. Huang et al. [37] analyzed open-domain DS from three main challenges (semantics, consistency, and interactivity). Guo et al. [32], Santhanam and Shaikh [86] interpreted the DS from the perspective of conditional text generation technology. Deriu et al. [16] surveyed the methods and concepts developed for the evaluation of dialogue systems.

Gao et al. [26] grouped conversational systems into three categories: (1) question answering agents, (2) task-oriented dialogue agents, and (3) social bots. Many later works [16, 18, 78] follow this classification. Figure 2 shows the research trend of these 3 kinds of DS in the past five years.

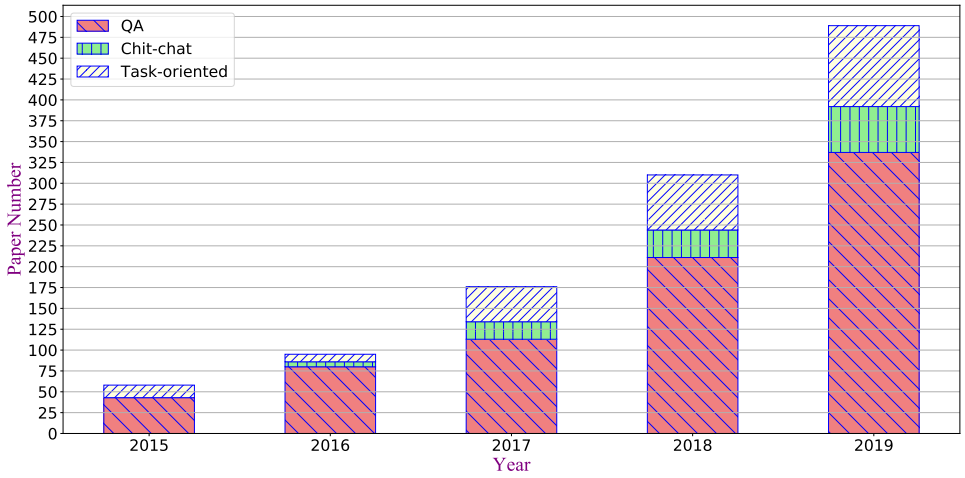

Fig. 2. The search results using keywords in paper titles on arxiv.org. "Question answering, Machine Reading Comprehension, QA, MRC", "chatbot, chat bot, chit chat, social bot, socialbot" and "task/goal-oriented, task/goal-completion, task/goal-driven" are used respectively.

There are obvious differences in dialogue patterns among the 3 types of DS. Dialogue state tracking (DST) and dialogue policy management (DPM) play an important role in the task-oriented DS which helps users accomplish tasks ranging from meeting scheduling to vacation planning. The dialogue focuses on filling predefined semantic slots with respect to special application scenarios.

The chit-chat bot can converse seamlessly and appropriately with humans about current events and popular topics, and often plays the role of chat companion or recommender. The goal of the system is to act as an emotional partner, and the main concern in the conversation is fluency and coherence. Generally, the more turns a conversation maintains, the more successful the role of an emotional partner becomes.

The QA systems can be categorized into answer selection and answer generation. The former is a retrieving or ranking problem between question and answer candidates, sometimes involving text as background. The later one is a task to select a span or generate an answer based-on the given document(s). The QA-based MRC technology has been greatly developed in the past few years. Equipped with rich knowledge drawn from various data sources including Web document(s) and pre-compiled knowledge graphs (KGâĂŹs), the QA agent is able to provide concise direct answers to user's queries. The main concern of the QA systems is the accuracy of the answers.

The DGDS tries to establish a real-world conversation mode as mentioned by former researchers [18, 26], where the interlocutors all ask and answer questions and provide each other with a mixture of facts and personal feelings during their general discussion. In other words, the DGDS maintains a dialog pattern in which relevant information can be obtained from the document(s). The similarities between the DGDS and the QA is that they both perform informative replies. The DGDS resembles the Chit-Chat in free form responses.

2.2 Differences Between the CRC and the DBD

The DGDS is composed of the CRC and the DBD. In the CRC, users put forward a series of questions for the given document(s), some words of these questions have mutual referential relations. Therefore, the bot needs to consider the history of dialogue to understand the current question. The CRC has gradually become the research hotspot [52] nowadays. The task of the CRC could be formulated as below: Given the document D, the conversation history with previous questions and answers $C = \{q_1, a_1, \ldots, q_{n-1}, a_{n-1}\}$ and the current question q_n, the goal is to predict the right answer a_n by maximizing the conditional probability $P(a_n|D, C, q_n)$[7]. The main difference between the CRC and the other MRC task is the dialogue history. In order to understand the current question accurately, agents need to solve the problems of coreference and ellipsis in the history of dialogue.

While in DBD, the conversation history is not QA pairs but utterances $U = \{u_1, u_2, \ldots, u_n\}$, the target is to predict u_{n+1} with $P(u_{n+1}|D, U)$. Therefore, the common ground of the CRC and the DBD is that they both need to jointly model document(s), historical conversations and last utterance, and they both need to pick out the required evidence from the document(s) and dialog history. For example, the Holl-E [66] dataset constructed by copying or modifying the existing information from a document which is relevant to the conversation, This setup is very similar to CoQA [81] where the answer response is extracted from a document as a span.

The distinctions between the CRC and the DBD lie in dialog patterns, evaluation methods, etc. The DBD takes advantage of document information to generate a reply based on understanding historical dialogue. The CRC tasks usually need to find out the location of the answer based on the understanding of the current question with the help of dialogue history. When encountering unanswerable questions, the CRC (e.g. QuAC) usually gives a "CANNOTANSWER", while DBD can response more freely, the reaction can be a direct answer as "I don't know", a rhetorical question or a changing subject. When evaluating the performance, the CRC tasks care about accuracy more, while the DBD tasks pay more attention to fluency, informative, coherent, etc.

Table 2. Differences among dialog systems. * means QA contains other task besides CRC and MRC, but we only compare these two here. In fact, the MRC includes the CRC, but we use MRC to represent single-turn QA in this table. # means we only compare with multi-turn chit-chat.

Characteristic		DGDS		QA*	Task-oriented	Chit-chat#
		DBD	CRC	MRC		
Function	Task-completion				✓	
	Emotional Partner					✓
	Providing Information	✓	✓	✓		
	Exchange Information	✓				
Knowledge	Utterance History	✓	✓			✓
	Background	✓	✓	✓	✓	
	Knowledgebase				✓	
	Document(s)-based	✓	✓	✓		
Evaluation	Fluency	✓				✓
	Coherence	✓				✓
	Informative	✓			✓	
	Diversity	✓				✓
	Turns	✓			✓	✓
	Accuracy		✓	✓	✓	

[7]Except in the ShARC, the question q_0 needed to be answered is asked by User in the beginning of the entire conversation, and the bot needs to ask the follow-up questions based on this q_0, then give an answer a_0 to close the dialog.

2.3 Summarization

To sum up, the DGDS is distinguished from other conversation scenarios in terms of dialog characteristic, the main concern, knowledge source, etc. In Table 2, we list the differences not only between the DGDS and the 3 DS categories, but also between the CRC, the DBD, and the MRC.

3 ARCHITECTURE

Most recently, a number of DGDS datasets [11, 29, 66, 73, 81, 83, 118, 138] and DGDS models [3, 8, 9, 31, 36, 39, 48, 55, 62, 68, 75, 76, 82, 91, 95, 100, 126, 127, 132, 135, 136, 141] have been proposed to mine unstructured document information in conversation. According to the characteristics of the task, current approaches normally consist of 5 parts: **joint modeling (JM)**, **knowledge selection (KS)**, **response generation (RG)**, **evaluation (EV)** and **memory (MM)**. We define the JM and the KS as the NLU problem and reckon the RG and the EV as the NLG problem. The general architecture of the DGDS is in Figure 3, we will introduce JM, KS, RG, and EV in this chapter. The memory (MM) module that the researchers haven't studied in depth in the DGDS untill now will be discussed in the Future Work chapter.

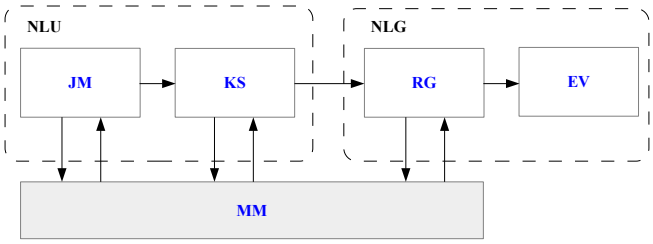

Fig. 3. The general architecture of the DGDS and the DBD.

3.1 Joint Modeling

Joint modeling refers to the integration of all input information. Joint modeling in the DGDS integrates the background document(s), current utterance and the dialogue history. Factors such as temporal relationship, connections among context, current utterance, and document(s), etc. need to be considered. When it comes to the differences in modeling, compared with the MRC model, the DGDS needs to consider conversation history, and compared with chit-chat, the DGDS needs to take background knowledge into account.

In the neural DS, Lowe et al. [58] first presented a method for incorporating unstructured external textual information for predicting the next utterance. Their model is an extension of the dual-encoder model [59]. There is a gap between the model they proposed and the actual application scenario. Yang et al. [123] incorporated external knowledge with pseudo-relevance feedback and QA correspondence knowledge distillation for a response ranking task. They built matrices on the word-level and semantic vector-level similarity. These methods treated the utterance history as a parallel relationship. Many DGDS models [3, 39, 62, 75, 76, 81, 82, 100, 132, 135, 141] followed this parallel setting.

Zhang et al. [133] weighted previous conversations with a turns-aware aggregation design. Zhou et al. [140] constructed the historical dialogue into 3-D tensor to extract the features between different rounds of dialogue and

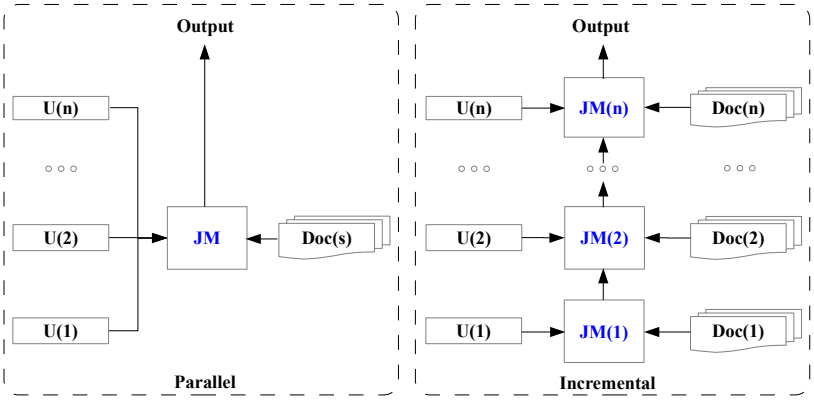

Fig. 4. The comparison of Parallel Modeling and Incremental Modeling. U is short for utterance, Doc is short for document.

response. These works treated the dialogue history differently according to temporal relationships, which we reckoned an incremental setting. Some DGDS [8, 31, 36, 48, 127] models employed the similar idea.

According to the way the previous DS model and the DGDS model handled historical conversations and text, we classify the JM into **Parallel Modeling** and **Incremental Modeling** as Figure 4. The parallel modeling means that the document(s), the historical dialogue and the last utterance are processed as a parallel relationship. The incremental modeling preserves the temporal relationship between the historical conversation and the current conversation and models them successively with the document(s).

3.2 Knowledge Selection

The knowledge selection (KS) in the DGDS is defined as the information seeking process in the document(s) and the utterance history that can be used to construct the response. According to whether there is an interpretable reasoning path when selecting, we divide the KS into **selection** and **reasoning**. In terms of the processed objects, we could divide the KS into **context-based** and **document(s)-based**.

The selection method in the DGDS normally extracts text segments [66, 73, 141] or selects some history utterances [75, 76]. While the reasoning method tries to build an interpretable reasoning path to the evidence in document(s) [8, 36].

The DGDS task parallels a growing interest in developing datasets that test specific reasoning abilities: algebraic reasoning [14], logical reasoning [113], commonsense reasoning [69] and multi-fact reasoning [41, 98, 111]. In MRC, Dua et al. [21] propose a task for discrete reasoning over the content of paragraphs. They define reasoning as subtraction, comparison, selection, addition, count, coreference resolution, etc. In the multi-turn DS, Wu et al. [116] employ a multi-hop attention with termination mechanism [92] for reasoning.

In the CRC, Saeidi et al. [83] reckoned reasoning ability as a fundamental challenge. An example is showed in table 3. According to the input content, the system needs to infer the missing conditions and put forward the corresponding questions to answer the initial question. Another example is given from CoQA in Table 5. The understanding of Question

4 and the reasoning of Answer 4 are based on the document(s) and the former QA pairs. In DBD, Liu et al. [55] performed reasoning on a knowledge augmented graph.

Table 3. One example in the ShARC dataset.

Speaker	Text Type	Text content
	Rule Text	YouâĂŹll carry on paying National Insurance for the first 52 weeks youâĂŹre abroad if youâĂŹre working for an employer outside the EEA.
User	Scenario	I am working for an employer in Canada.
User	Target Question	Do I need to carry on paying UK National Insurance?
Bot	Follow-up Question	Have you been working abroad 52 weeks or less?
User	Follow-up Answer	Yes
Bot	Target Answer	Yes

3.3 Response Generation

In the DGDS, there are three ways to generate replies: extraction [3], retrieval [135], and generation [48]. Extraction needs to determine the beginning and end of the text, retrieval needs to select one from the candidates the highest score among the given candidates, and generation needs to generate words in turn to form a complete reply. Based on whether generating all words, we consider the extraction and retrieval to be **indirect**, while generation to be **direct**. Some works try to combine the indirect and direct are defined as **hybrid** [62, 136].

Language models [4, 64] can predict the next word given the sequence of the previous word, so they are widely used in the task of text generation. For example, encoder-decoder model [10], sequence-to-sequence model [97] and attention based models [17, 72, 77, 105, 124] greatly improve the text generation tasks. We observed that the attention-based generation are often adopted in the DGDS.

3.4 Evaluation

EV is used to judge whether the generated text meets the DGDS requirements. In the CRC, we normally adopt to some mature metrics of information retrieval (IR) such as accuracy. The DBD models are usually measured with the word overlap metrics (F1, BLEU, etc), which are insufficient for dialog scenario [57, 101], hence human evaluations are usually employed paralleled with auto evaluations. We still need to establish an auto evaluation metric highly correlated with human evaluations for dialog quality measuring.

Generally speaking, DS with good performance needs to have the characteristics of high semantic relevance, rich information, and diverse expressions. There are many works addressed the evaluation methods of the DS. Paek [70] studied what purpose dialogue measurement serves, and then propose an empirical method to evaluate the system that meets that purpose. Liu et al. [50] introduced embedding-based relevance evaluation metrics (the Greedy Matching, the Embedding Average, The Vector Extreme). Lowe et al. [57] introduced ADEM for mimic human evaluation. Kannan and Vinyals [40] discussed the adversarial evaluation in the DS. Xu et al. [119] proposed mean diversity score (MDS) and probabilistic diversity score (PDS) to evaluate the diversity of responses generated when multiple reference responses are given. Tao et al. [101] evaluated a reply by taking into consideration both a ground-truth reply and a query. Sai et al.

[85] pointed out the drawback of the ADEM and outlined we still have a long way to go in the automatic evaluation of DS. Similar to other DS, the DGDS is also in an era of lack of automatic evaluation indicators.

3.5 Summarization

At present, the models used to solve the CRC problems are mainly extractive and generative, while retrieval and generative models are usually adopted to address the DBD. We summarize the differences between the CRC and the DBD in dealing with the 4 model components in Table 4.

Table 4. Main architecture in the DGDS. "Acc." means accuracy.

architecture	CRC		DBD	
	Extractive	Generative	Retrieval	Generative
JM	Parallel / Incremental	Parallel	Parallel	Parallel / Incremental
KS	Selection	Reasoning	Selection	Selection / Reasoning
RG	Indirect	Direct / Hybrid	Indirect	Direct / Hybrid
EV	Acc./ Word overlap	Acc./ Word overlap	Acc.	Lack of standards*

4 DATASETS

We introduce the DGDS related datasets of the CRC and the DBD respectively.

4.1 CRC Datasets

Recently, a series of CRC datasets are proposed by researchers. For instance, Reddy et al. [81] released CoQA, a dataset with 8,000 conversations about given passages from seven different domains. We present a dialogue example in the CoQA in Table 5. The two interlocutors conduct a question and answer dialogue according to the given text.

Table 5. The CoQA dialogue example. The same text color corresponds to the QA pair and the evidence.

Passage:	Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and MelanieâĂŹs husband Josh were coming as well.		
Question 1:	Who had a birthday?	Answer 1 :	Jessica
Question 2:	How old would she be?	Answer 2 :	80
Question 3:	Did she plan to have any visitors?	Answer 3 :	Yes
Question 4:	How many?	Answer 4 :	Three
Question 5:	Who?	Answer 5 :	Annie, Melanie and Josh

At the same time, Choi et al. [11] introduced QuAC, Compared with the CoQA, document(s) are only given to the answerer, whereas the questioner asks the questions based on the title of passages. The answerer replies to the question with a subsequence of the original passage and determines whether the questioner can ask a follow-up question. Hence the dialogs often switch topics compared with CoQA's dialogs including more queries for details. The CoQA answers are less than 3 tokens long on average, while QuACâĂŹs are over 14. Both CoQA and QuAC have a question type prediction subtask. The training set of the QuAC has one reference answer, while dev and test set questions have

multiple references each. We present an instance of the dev set of QuAC in Table 6. We only keep the reference answers with obvious differences, Compared with the DBD's multiple references example in Table 7, the diversity of the CRC's multiple references is insufficient. This is suffered from extracting fragments from the document(s) as reference.

Table 6. The multiple reference example of the QuAC dev set.

Category	Text
Background	"Anna Vissi ... also known as Anna Vishy, is a Greek Cypriot singer, ..."
Document	"In May 1983, she married Nikos Karvelas, a composer, with whom she ... "
question 1/2/3	"what happened in 1983?" / "did they have any children?" / "did she have any other children?"
question 4 answers	"what collaborations did she do with Nikos?" "After their marriage, she started a close collaboration with Karvelas ..." "a composer, with whom she collaborated in 1975" "Thelo Na Gino Star" "Since 1975, all her releases have become gold or platinum and ..."
question 5/6/7	"what influences does he have ..." / "what were some of the songs?" / "how famous was it?"
question 8 answers	"did she have any other famous songs?" "In 1986 I Epomeni Kinisi (The Next Move) was released." "Epomeni Kinisi (The Next Move) was released." "Pragmata (Things)" "The album included the hit Pragmata (Things) and went platinum,"

Sequence QA attracted attention in recent years. The RC2 dataset [118] leverages knowledge from reviews to answer multi-turn questions from customers, which is an open-domain CRC task. The QBLink [22] is a sequential QA dataset asking multiple related questions about a Wikipedia page. The questions are designed primarily to challenge human players in Quiz Bowl tournaments. However, due to the continuous change of the document(s) for the question, they cannot be classified into the DGDS. Ma et al. [61] also introduced a CRC dataset. We did not include it here for that it both treats dialogues as documents [96] and limits task to entity completion. Unlike other CRC datasets where conversations are led by the users, Saeidi et al. [83] introduced SHARC, a sequential QA task based on regulation texts given by government websites and the bot lead the dialogue interaction. When users ask a question about laws and regulations based on a certain scenario, the bot will give accurate answers through a series of queries to users. The bot needs to understand complex decision rules based on the conversation history and certain scenarios to answer the initial question raised by users, while other CRC tasks usually extract answers directly from texts. Besides, the SHARC task needs to develop free-form follow-up questions to determine whether the user meets the decision rules, while other CRC tasks do not.

Iyyer et al. [38] initiated sequential QA on a structured knowledge table. Saha et al. [84] and Guo et al. [33] both introduced the task of Complex Sequential QA based on KG. Most recently, Christmann et al. [12] proposed ConvQuestions datasets which are also based on a KG. These sequential QA tasks based on structured data are not included in the DGDS. There are other interesting tasks such as constructing QA pairs from a document [43], asking right sequential questions about a document [24], decomposing complex questions into a sequence of simple questions [98], which have some potential connections with the CRC.

Table 7. The one-to-many examples in Holl-E test set. The movie name is "The Secret Life of Pets", we do not present the documents (comments, review, plot) for saving space.

Speaker	Reference	Utterance
User		What do you think about the movie?
Bot		I think it was comical and entertaining.
User		It delivered what was promised.
Bot	A	I agree! IâĂŹm surprised this film got such a low overall score by users.
	B	My favorite character was Gidget! She was so much fun and so loyal to her friends!
	C	Yes! As a Great Dane owner, I often wonder what my dogs are thinking. It was fun to see this take on it.
	D	It was full of cliches with a predictable story, but with some really funny moments.

4.2 DBD Datasets

Recently, a number of DBD datasets based on movie domain have been released [66, 138]. The document(s) they discussed come from multi-sources (e.g. Wikipedia), the conversations they used are often collected from crowdsourcing platform Amazon Mechanical Turk (AMT) or Reddit.

The CMUDoG [138] is a dataset built with AMT where dialog is generated based on the given background text. The Holl-E dataset [66] also addresses the lack of background knowledge in the DS. The difference between CMUDoG and Holl-E are list below:

- The CMUDoG only uses Wikipedia article, mainly the plot of the movie, which is divided into four sections, one section is the basic information of a movie such as reviewers' comments, introductions, ratings, and the other three sections are the plots of the movie). The Holl-E uses Reddit Comments, IMDB and Wikipedia introductions as the background.
- The CMUDoG is discussed according to the sections. While the Holl-E can discuss any part of the given document(s) as long as it is relevant to the current conversations, and the author gives the source evidence of each utterance in the dataset.
- The CMUDoG has two scenarios (whether the interlocutor knows the document(s) or not), while both Holl-E's interlocutors are aware of the content of the document(s).
- There are two versions of the test set in the Holl-E: one with a single golden reference and the other with multiple golden references, while the CMUDoG only has one test set with one single golden reference for each dialogue. We present a dialog example of Holl-E in Figure 7.
- The Holl-E collects self-conversations [42] which means the same worker plays the role of both parties in the conversation.
- The Holl-E has five manual ratings for conversational fluency similar to the CoQA, while CMUDoG uses BLEU and dialog turns to classify the dialogues. The BLEU here measures the overlap of the turns of the conversation with the sections of the document.
- The dialogues in the Holl-E allow one speaker to freely organize the language, the second speaker needs to copy or modify the existing information, also similar to the CoQA. But unlike the CoQA, the second speaker can properly add words before or after span to ensure smooth dialogue. To comparison, the CMUDoG placed no limits on dialogues except the two interlocutors respectively play an implicit recommender and recommended role.

To provide the conversation model with relevant long-form text on the fly as a source of external knowledge, Qin et al. [73] presented a dataset where the next utterance is generated with web-text. Given the dialogue history obtained from Reddit, the web text is treated as an external source to generate the next dialogue related to the dialogue history and containing external knowledge. Some symbols on the original web page file are retained as annotations, such as "#" annotating the dialogue, the corresponding keyword or paragraph in the document, "<title>" annotating the position of the title, "<p>" annotating the text structure, etc.

Topical-Chat [29] relies on multiple data sources, including Washington Post articles, Reddit fun facts and Wikipedia articles about pre-selected entities, to enable interactions where the interlocutors have no explicit roles. The external knowledge provided to interlocutors could be the same or not, leading to more diverse conversations.

Reddit conversational dataset is released by Dialog System Technology Challenges 7 (DSTC-7)[8] and is extracted from the Reddit website. For each web page of the Reddit website, there is a link below the title, which might provide background knowledge for the current topic. The data can be obtained from Reddit dump[9] and Common Crawl[10]. After further filtering, it may be used in the DGDS, but it does not meet our definition at present because the background knowledge is either redundancy or has low relevance with the conversation.

There are some other works trying to incorporate unstructured documents knowledge into DS. Vougiouklis et al. [106] proposed a dataset aligning knowledge from Wikipedia in the form of sentences with sequences of Reddit utterances. Dinan et al. [18] released a dataset which constructed with ParlAI [65] to provide a supervised learning benchmark task which exhibits knowledgable open dialogue with clear grounding. However, these datasets remove the organization of the document(s) by flattening all the paragraphs into separate sentences. Akasaki and Kaji [2] proposed a task to initiate a conversation based on the given document, which has some potential connection with the DBD.

4.3 Summarization

Table 8 summarize the characteristics of above mentioned DGDS datasets in 6 aspects: chat style, domain of dialogue, whether the speaker can see the document(s), whether the dataset is labeled for training, which side leads the conversation.

Table 8. Comparison of Document Grounded Conversation Datasets. "Flu." is short for Fluency.

Dataset	QA	See Doc	Dialog.Source	Domain	Labeled	Lead by
CoQA[81]	✓	Both	AMT	Open	Span	User
QuAC[11]	✓	Bot	AMT	Open	Span	User
ShARC[83]	✓	Both	AMT	Regulatory	No	Bot
RC2[118]	✓	Both	Manually	Review	Span	User
CMUDoG[138]		User/both	AMT	Movie	No	Both
Holl-E[66]		Both	AMT	Movie	Flu./Span	Both
CbR[73]		All	Reddit	Open	No	Multi
T-Chat[29]		Both	ParlAI/AMT	Open	No	Both

Table 9 illustrates the statistics of the DGDS datasets with total dialog numbers, turns per dialogue, total number of document(s), average words of document, average word of each utterance. It is worth mentioning that the definition of

[8]https://github.com/mgalley/DSTC7-End-to-End-Conversation-Modeling/tree/master/data_extraction
[9]http://files.pushshift.io/reddit/comments/
[10]http://commoncrawl.org/

"Turns" represents a QA pair in the CRC and one utterance in the DBD. We can observe that the DBD datasets have more dialog turns, more words in both document(s) and utterances.

Table 9. Statistics of Document Grounded Conversation Datasets. The statistics with the asterisk are from us.

dataset	Dialogs	Turns/Dialog	Num.of.Doc.	Words/Doc.	Words/Utter.
coQA[81]	8,399	15.2	8,399	271	4.1
QuAC[11]	13,594	7.2	8,854	401	10.6
ShARC[83]	32,436	2.7*	948	60.6*	5.2*
RC2[118]	1,218*	3.9*	1,218*	108.1*	4.3*
CMUDoG[138]	4,112	21.4	120	229	18.6
Holl-E[66]	9,071	10.0	921	727.8	15.3
CbR[73]	2.82M	86.2	32.7k	7,347.4	18.7
T-Chat[29]	11,319	21.9	3064*	830*	19.7

5 CRC MODELS

In this chapter, we summarize and analyze the current CRC models according to the architecture defined previously. The existence of the historical answers of the CRC entails differences in JM and KS with the DBD. For example, in the KS processing of the CoQA, the corresponding answer position can be predicted for each historical question, which means that in the JM part of the CRC, the historical answers are not necessarily added to the model, but only used as a midterm target for the RS ability training.

5.1 Joint Modeling in the CRC

In CRC, joint modeling (JM) aims to integrate the background document(s) and dialogue history (QA pairs) in Table 4.

5.1.1 Parallel Modeling. We further classify the parallel modeling into complete parallel and partial parallel depends on whether concatenating all utterances together.

Complete Parallel. Reddy et al. [81] proposed a hybrid model, DrQA [6] + PGNet [30, 88], which combines the sequence-to-sequence model and MRC model together to extract and generate answers. To integrate information of conversational history, they treated previous question-answer pairs as sequence and append them to the document. Similar to Reddy et al. [81], Zhu et al. [141] proposed SDNet which appends previous question-answer pairs to the current question, while in order to find out the related conversational history, they employed additional self-attention on previous questions.

Su et al. [95] defined two types of questions, verification ones and knowledge-seeking ones in the CRC and proposed an adaptive framework for the CoQA. They first extracted the rationale for the question from the document(s), then used different components for the corresponding question type. Ju et al. [39] used RoBERTa[54] combining with adversarial training (AT) [28] and knowledge distillation (KD) [25] to leverage additional training signals from well-trained models. They used parallel encoding, where $Q_k^* = \{Q_1, A_1, \ldots, Q_{k-1}, A_{k-1}, Q_k\}$, and concatenated Q_k^* with document(s) as the input of RoBERTa. Since the answers of CoQA dataset can be free-form text, Yes, No or Unknown, and besides the Unknown answers, each answer has its rationale, an extractive span in the passage, they also add a new task named Rationale Tagging in which the model predicts whether each token of the paragraph should be included in the rationale.

In other words, tokens in the rationale will be labeled 1 and others will be labeled 0. For unknown questions, they should all be 0.

On the ShARC task, Zhong and Zettlemoyer [136] proposed an Entailment-driven Extracting and Editing (E3) model, Sharma et al. [91] presented an UrcaNet model to learn the deep level clues, Lawrence et al. [46] made the sequence generation process bidirectional by employing special placeholder tokens. They all adopted the parallel modeling which treated the rule text, scenario and context equally.

Partial Parallel. Yatskar [126] used BiDAF++ [13] with ELMo [72] to answer the question based on the given document and conversational history. Besides encoding previous dialog information to the context representation, they labeled answers to previous questions in the context. They proposed to first make a Yes/No decision, then output an answer span only if Yes/No was not selected.

Ohsugi et al. [68] proposed a fine-tuning BERT (w/k-ctx)[11] model, which treated questions and answers as independent input, concatenated them to a passage and encode with Bert. Qu et al. [76] proposed a general framework for the QuAC in an information-seeking point of view. Some other works only verified on the CoQA development set.

5.1.2 Incremental Modeling. Huang et al. [36] adopted a flow mechanism to better understand conversational history. The FlowQA model employed an alternating parallel processing structure and incorporated intermediate representations generated during the process of answering previous questions. We classify the FlowQA into the incremental modeling as the encoding information of historical dialogue is accumulated from far to near turn. Yeh and Chen [127] further considered the long-distance historical dialogue information in the process of reasoning. They also proposed a Bert-FlowDelta to investigate the extension of Bert into the multi-turn dialog. These models carried out experiments on both the CoQA and the QuAC.

Chen et al. [8] proposed a GraphFLOW structure, which built a sparse graph from the document and utterances history dynamically then reasoned with the Graph-Flow mechanism which sequentially processed the graphs they constructed. Following Choi et al. [11], they leveraged conversation history by concatenating a feature vector encoding previous N answer locations to the document(s) word embeddings. They also prepended the previous N QA pairs to the current question and concatenated a 3 dimension relative turn marker embedding to each word vector in the augmented question to indicate which turn it belongs to.

Experimenting on the CoQA development set, Gu et al. [31] proposed a TT-Net model on CoQA, which was capable of capturing topic transfer features using the temporal convolutional network (TCN) in the dialog. The TT-Block packaged by the BiLSTM, TCN and Self-attention mechanism was presented to extract topic transfer features between questions, which was also an incremental modeling method.

5.1.3 Summarization. The way of parallel modeling is more conducive to the screening of historical dialogue information, and the way of incremental modeling is more in line with the logical order of reasoning. Task characteristics and design ideas determine the modeling method. In table 10 we summarize some modeling methods of combining document(s), last question and context in the CRC.

5.2 Knowledge Selection in the CRC

As we defined in the architecture section, the KS in the CRC task can be divided into context-based and document(s)-based.

[11](k-ctx) means employing previous k QA pairs.

Table 10. The modeling methods in the CRC. C means $\{Q_1, A_1, Q_2, A_2, \ldots, Q_{n-1}, A_{n-1}\}$, i={1, 2, …, n-1}.

Description	Method	Examples
Q and A sequentially	$\{C, Q_n\}, \{Doc\}$	[39, 81, 95, 141]
Doc, Q and A separately	$\{Doc\}, \{Q_n\}, \{Q_i\}, \{A_i\}$	[11, 36, 127]
Doc and Q separately	$\{Doc\}, \{Q_i\}$	[8, 31]
Concatenating all	$\{C, Q_n, Doc\}$	[46, 75, 83, 91, 136]
Doc paired with Q and A	$\{Q_n, Doc\}, \{Q_i, Doc\}, \{A_i, Doc\}$	[68]
Doc, latest Q and $\{Q_i, A_i\}$	$\{Doc, Q_n, Q_i, A_i\}$	[76]

5.2.1 Context-based. Qu et al. [75] introduced a general framework containing an utterances history selection module that retrieves a subset of dialog history more useful than others[12]. They also presented a history answer embedding module to incorporate the utterance history naturally to BERT. Qu et al. [76] further improved previous work [75] with three aspects. Firstly, a positional history answer embedding (PosHAE) method based on location information was proposed. Secondly, a history attention mechanism (HAM) was used to "soft select" the utterance history. Thirdly, in addition to dealing with conversation history, multi-task learning (MTL) was used to predict the dialog act.

5.2.2 Document(s)-based. In this section, we introduce the current CRC model of the KS for the document(s) from implicit and explicit reasoning perspectives.

Implicit Reasoning. As mentioned earlier, the FlowQA proposed by Huang et al. [36] was an implicit reasoning module which uses historical dialogues to perform information flow in the document(s). The problem is that the learned representations captured by the FLOW change during multi-turn questions. Whether such changes correlated well with the current answer or not is unclear. To explicitly explore the information gain in FLOW and further relate the current answer to the corresponding context, Yeh and Chen [127] presented FlowDelta, which focused on modeling the difference between the learned context representations in multi-turn dialogues. They add a subtraction of the previous hidden state into the flow section to mimic the conversation topic change between sequential turn questions. To capture the topic transfer (TT) information in CoQA, Gu et al. [31] proposed a TT-Net model using TT-Block packaged by the BiLSTM, TCN, and Self-attention mechanism is presented to extract topic transfer features between questions.

Explicitly Reasoning. Chen et al. [8] argued that the Integration-Flow mechanism in FlowQA fails to mimic the human reasoning process since humans do not first perform reasoning in parallel for each question, and then refined the reasoning results across different turns. The reasoning performance at each turn was only slightly improved by the hidden states of the previous reasoning process. So they built a sparse graph from the document and utterances history dynamically, then proposed a GraphFLOW (GF) structure to perform interpretable reasoning. The GF mechanism adopted a GNN to the sparse graph and fused both the original context information and the updated context information at the previous turn. The final prediction results were obtained through multiple rounds of GF operation on dialogues. On the ShARC task which takes reasoning ability as the main challenge, the E3 model [136] jointly extracted a set of decision rules from the procedural text while reasoning about which was entailed by the conversational history and which still needed to be edited to create questions for the user. To prevent models from learning the superficial clues, Sharma et al. [91] modified the ShARC dataset by automatically generating new instances reducing the occurrences of those patterns. They proved that Zhong and Zettlemoyer [136] relied more heavily on spurious clues in the dataset and suffered a steeper drop in performance on the ShARC-augmented dataset.

[12] Alough they simply select the nearest three utterances.

5.3 Response Generation in the CRC

In the CRC, the indirect generation usually means determining the probability of each word as the starting and ending position of the answer. While in the direct generation, the decoder of the generative model generates the response word by word.

5.3.1 Indirect Generation. On the CoQA and the QuAC dataset, Reddy et al. [81] and Choi et al. [11] use DrQA and BiDAF++ for answer span prediction respectively. Models on CoQA and QuAC [8, 31, 36, 39, 68, 75, 76, 126, 127, 141] all followed the span prediction setting.

5.3.2 Direct Generation. Reddy et al. [81] proposed a Pointer-Generator network (PGNet) model generating the answer using an RNN decoder which attends to the encoder states. They also employ a copy mechanism in the decoder which allows copying a word from the document. Su et al. [95] adopted a seq2seq model as the DistillNet to refine the final answer. On the ShARC dataset, the questions to generate are derived from the background text. Saeidi et al. [83] presented a Combined Model (CM) which is a pipeline that contains a rule-based follow-up question generation model. Lawrence et al. [46] proposed bidirectional sequence generation model and introduced several different sequence generation strategies.

5.3.3 Hybrid Generation. Zhong and Zettlemoyer [136] first selected a rule-span, then used a separate attentive decoder to generate the pre-span and the post-span words, the concatenation of the three parts constituted the final output.

5.4 Evaluation Method in the CRC

Since the CRC answers are usually a clear span or "YES/NO/UNANSWERABLE", the evaluation metrics are usually automatic. Besides EM (exact match), former work [8, 11, 39] employed a word-level $F1$ similar to Rajpurkar et al. [79] to measure the performance of model. This metric regards the prediction and ground truth as bags of tokens and takes the maximum F1 over all of the ground truth answers for a given question, and then averages over all of the questions.

Choi et al. [11] proposed human equivalence score (HEQ). The HEQ-Q is the accuracy of each question, where the answer is considered correct when the modelâĂŹs F1 score is higher than the average human F1 score. Similarly, the HEQ-D is the accuracy of each dialog âĂŞ it is considered correct if all the questions in the dialog satisfy the HEQ. A system that scores the value of 100 on the HEQ-D can by definition achieves average human performance over full dialogs. Chen et al. [8], Huang et al. [36], Ohsugi et al. [68] also used this metric.

On the ShARC task, for all following classification tasks, Saeidi et al. [83] used micro- and macro- averaged accuracy. For the follow-up generation task, they computed the BLEU scores between the reference and generated follow-up questions. Then Zhong and Zettlemoyer [136] showed a combined metric ("Comb."), which was the product between the macro-averaged accuracy and the BLEU-4 score on ShARC dataset.

We list the current model performance in the CRC in Table 11. We only show the single models which are published on the QuAC[13] and the CoQA[14] leaderboard. BERT + Answer Verification[15] and XLNet [124] + Augmentation[16] are added to the table[17] for comprison. We also present the model performance on the ShARC[18] dataset.

[13] http://quac.ai/
[14] https://stanfordnlp.github.io/coqa/
[15] https://github.com/sogou/SMRCToolkit
[16] https://github.com/stevezheng23/xlnet_extension_tf
[17] The top single model performance of the QuAC leaderboard is F1=73.5, HEQ-Q=69.8, HEQ-D=12.1 untill Beijing time, November 29, 2019 0:00.
[18] https://sharc-data.github.io/

Table 11. The model performance on the CoQA, QuAC, and ShARC Test set. * means less training time.

Model	CoQA	QuAC			ShARC	
	F1	F1	HEQ-Q	HEQ-D	(Micro/Macro)-Acc.	BLEU-(1/4)
DrQA+PGNet [81]	65.1	-	-	-		
BiDAF++ w/ 2-ctx [11]	-	60.1	54.8	4.0		
Bert+HAE* [75]	-	62.4	57.8	5.1		
FlowQA [36]	75.0	64.1	59.6	5.8		
HAM [76]	-	65.4	61.8	6.7		
SDNet [141]	76.6	-	-	-		
GraphFlow [8]	77.3	64.9	60.3	5.1		
Bert-FlowDelta [127]	77.7	67.8	63.6	12.1		
Bert w/ 2-ctx [68]	78.7	64.9	60.2	6.1		
BERT + Ans.Verification	82.8	-	-	-		
XLNet + Augmentation	89.0	-	-	-		
RoBerta + AT + KD [39]	90.4	-	-	-		
CM [83]					0.619 / 0.689	54.4 / 34.4
E3 [136]					0.676 / 0.733	54.1 / 38.7
BiSon [46]					0.669 / 0.716	58.8 / 44.3
UrcaNet [91]					0.659 / 0.717	61.2 / 45.8
Human	88.8	81.1	100	100		

6 DBD MODELS

In this chapter, we analyze the current DBD models in accordance with the architecture previously presented.

6.1 Joint Modeling in the DBD

Same as the CRC chapter, we introduce the JM of DBD from parallel and incremental as well.

6.1.1 Parallel Modeling. Zhao et al. [135] focused on connecting dialogue and background knowledge and identifying document information for matching. They used a retrieve-based method DGMN. By hierarchical interacting with the response, the importance of the different parts of the document and context was dynamically determined in the matching step. They conducted their experiments on the CMUDoG and Persona-chat [129] as retrieval tasks. Arora et al. [3] modeled the latest utterance with prober and responder history respectively, then used the enhanced last utterance to filter the document segments. Qin et al. [73] adopted all utterance history as query to retrieve related text segments from document(s). Tang and Hu [100] also employed a parallel way of encoding the text. They proposed three models: DialogTransformer which consisted of knowledge memory and response generation, DialogTransformer-Plus which used a three-stage multi-head attention mechanism to incorporate dialogue utterances and knowledge representation, and DialogTransformerX which combined different sources in the generation. Models [82, 132] focusing on KS all employed the parallel modeling methods.

6.1.2 Incremental Modeling. Incremental means adding historical conversation, document and current utterance layer by layer. Li et al. [48] proposed Incremental Transformer with Deliberation Decoder (ITEDD). They designed an incremental transformer to encode multi-turn utterances along with knowledge in the related document(s). One problem of ITEDD is the authors only experiment on the CMUDoG where the document(s) is pre-defined to sections,

they did not consider the effect when the doc length increases and the deliberation decoder could not mine enough document(s) information.

6.1.3 Summarization. The advantage of incremental modeling is that it can reflect the temporal relationship of dialogue history, but it also loses the ability to filter dialogue history since long-distance dialogue information will be forgotten in encoding process, so the incremental modeling usually adopts the near rounds of dialogue, increasing the rounds of historical dialogue does not improve performance. In contrast, parallel modeling can better filter historical conversation information, but it needs to consider the influence of temporal relationship. In table 12 we sum up the modeling categories in DBD.

Table 12. The categories of modeling in DBD. D is Document, C=$\{U_1, \ldots, U_{n-1}\}$ is the utterances history, Bot=$\{Bot_1, \ldots, Bot_{n-1}\}$ is the utterances from Bot, User=$\{User_1, \ldots, User_{n-1}\}$ is the utterances from User, i={1, 2, …, n-1}.

Description	Method	Examples
Concatenating all	$\{D, C, U_n\}$	[138]
Ds and Us seperatly	$\{D\}, \{C, U_n\}$	[55, 62, 73, 82, 132, 135]
User and Bot seperatly	$\{D\}, \{User_n\}, \{Bot\}, \{User\}$	[3]
D, U, C seperatly	$\{D\}, \{U_n\}, \{C\}$	[100]
D, U, U_i seperatly	$\{D\}, \{U_n\}, \{U_i\}$	[48]

6.2 Knowledge Selection in the DBD

Unlike the CRC, the KS in the DBD usually pays little attention on context while focus on document(s). We divide the KS of the DBD into selection and reasoning.

6.2.1 Selection. In DGDS, Qin et al. [73] presented an architecture combining MRC technique [53] into next utterance generation with external web-text. They pre-selected some candidate segments for encoding which improves training efficiency. To improve the process of using background knowledge, Zhang et al. [132] focused on generation-based methods and propose Context-aware Knowledge pre-selection (CaKe). The main contribution of CaKe is also a knowledge pre-selection module. They firstly adopted the encoder state of the utterance history context as a query to select the most relevant knowledge, then employed a modified version of BiDAF to point out the most relevant token positions of the background sequence. A limitation of CaKe is that the performance decreases when the background document becomes longer.

Ren et al. [82] argued that previous KS mechanisms have used a local perspective, i.e., choosing a token at a time based solely on the current decoding state. They adopted a global perspective, i.e., pre-selecting some text fragments from the background knowledge that could help determine the topic of the next response. Their model firstly learned a topic transition vector to encode the most likely text fragments to be used in the next response, then used the vector to guide the local KS at each decoding timestamp.

Models focusing on evidence selection [62, 73, 82, 132] and concentrating on feature extraction [3] have no explicit reasoning path, they are all belong to the selection method.

6.2.2 Reasoning. In the DGDS, Liu et al. [55] claimed the first to unify knowledge triples and long texts as a graph. Then employed a reinforce learning process in the flexible multi-hop knowledge graph reasoning process, called AKGCM. They conducted experiments on Holl-E and WoW. One deficiency is that the reinforcement learning policy is doubtable

since they chose a labeled state as a reward. They used the top 1 accuracy to evaluate the performance of knowledge selection.

Approaches [48, 100] gradually integrating historical information during incremental modeling are also considered to be reasoning because of the existence of the implicit inference path.

6.3 Response Generation in the DBD

We analyze the RG from 3 categories: the direct generation that generates the response word by word, the indirect generation that predicts a span or ranks candidates, and the hybrid generation that combines indirect and direct.

6.3.1 Direct. In order to improve the consistency of context and the correctness of knowledge, the ITEDD model [48] employed a two-way deliberation decoder [117] for response generation. The first-level decoder took the representation of last utterance and last encoding state as input to generate responses contextual coherently. The second-level decoder took the first-level decoding results and the grounded document as input to guide the generation. The CMR model [73] chose an attentional recurrent neural network decoder after a memory that summarized the salient information from both context and document. The DialogTransformer model [100] used a one-layer Transformer as decoder.

6.3.2 Indirect. Zhao et al. [135] presented a retrieval-based model that fused information in the document(s) and context into representations of each other. They dynamically determined whether the grounding was necessary, and weighted the importance of different parts of the document and context through hierarchical interaction with a response at the matching step. Arora et al. [3] regarded Holl-E as a span prediction task, they argued that models computing rich representations for the document(s) and utterance suffered space and time when dealing with long text. Hence they adopted a knowledge distillation method to train a simple model that learned to mimic certain characteristics of a complex span prediction teacher model.

6.3.3 Hybrid. Qin et al. [73] firstly selected the document segment(s) then perform generation. The DialogTransformerX model [100] combines three methods for DG: (1) Generating a word, (2) Copying a word from dialogue utterance, (3) Copying a word from unstructured external knowledge. Combining the advantages of generative and extractive models. Meng et al. [62] propose RefNet model. At the decoding step, a decoding switcher predicts the probabilities of executing the reference decoding or generation decoding. Reference decoder learns to directly select a semantic unit (e.g., a span containing complete semantic information) from the background while generation decoder predicts words one by one. One disadvantage of the RefNet is that it needs labeled data for span prediction, hence it is not a general model for DGDS. Another disadvantage is that the emphasis of the RefNet is on the decoding part, not much on the joint modeling part where the importance of different utterances history should be considered.

6.4 Evaluation Method in the DBD

At present, the evaluation indexes of the DGDS can be divided into two categories: one is manual evaluation, the other is objective auto evaluation.

6.4.1 Human Evaluation. We divide human evaluations into the grading method and the comparison method.

Grading. Workers were asked to rate the responses generated by the model. Moghe et al. [66] let workers rate the response on a scale of 1 to 5 (with 1 being the worst) on the following four metrics: (1) Fluency, (2) appropriateness/relevance of the response under the current context, (3) humanness of the response, i.e., whether the responses look as if they were generated by a human (4) and specificity of the response, i.e., whether the model produced movie-specific

responses or generic responses such as âĂIJThis movie is amazingâĂİ. Meng et al. [62], Ren et al. [82] had workers annotate whether the response was good in terms of four aspects: (1) Naturalness (N), i.e., whether the responses are conversational, natural and fluent; (2) Informativeness (I), i.e., whether the responses use some background information; (3) Appropriateness (A), i.e., whether the responses are appropriate/relevant to the given context; and (4) Humanness (H), i.e., whether the responses look like they are written by a human. Li et al. [48] defined three metrics - fluency, knowledge relevance [51] and context coherence. All these metrics are scored 0/1/2. Fluency is whether the response is natural and fluent. Knowledge relevance is whether the response uses relevant and correct knowledge. Context coherence is whether the response is coherent with the context and guides the following utterances. Human Judgment employed by Tang and Hu [100] was to rate the response on a scale of 1 to 5 on the Fluency and Knowledge Capacity of the generated response. Gopalakrishnan et al. [29] asked two humans to separately annotate (possible values in parentheses) whether the response is comprehensible (0/1), on-topic (0/1), and interesting (0/1). They also asked workers to annotate how effectively the knowledge is used in response (0-3) and if they would like to continue the conversation after the generated response (0/1).

Comparison. Workers were required to directly compare the generation response of the model with the generation of the reference target or other baseline models, and choose the preferred one. In human evaluation, Liu et al. [55], Qin et al. [73] tested the preferences between the response from their model and comparator model. Outputs from systems to be compared were presented pairwise to judges from a crowdsourcing service.

6.4.2 Retrieval Auto Ealuation. The ranking of candidate answers is the core of the retrieval DS. Zhao et al. [135] used R@N which evaluates whether the correct candidate is retrieved in the top N results. Liu et al. [55] adopted a Hit@1 for the accuracy of selecting the right knowledge.

6.4.3 Generative Auto Evaluation. In this section, we introduce some automatic evaluation metrics currently used in the DGDS.

Perplexity [4]. In the language model, the Perplexity (PPL) is usually employed to measure the probability of the occurrence of a sentence. While in the DS, the PPL measures how well the model predicts a response, a lower perplexity score indicates better generation performance. The disadvantage of the PPL is that it can not evaluate the relevance between the response and the context.

Entropy [131]. Entropy reflects how evenly the empirical n-gram distribution is for a given sentence.

Distinct [47]. Distinct measures the diversity of reply by calculating the proportion of 1&2-grams in the total number of generated words to solve the problem of universal reply in DS.

F1 [79]. Word-level $F1$ treats the prediction and ground truth as bags of tokens and measures the average overlap between the prediction and ground truth answer.

BLEU [71]. Measuring the similarity by calculating the geometric average of the accuracy of n-gram between the generated response and the golden response.

ROUGE [49]. ROUGE is based on the calculation of the recall rate of the longest common subsequence of generating response and the real one.

METEOR [45]. In order to improve BLEU, METEOR further considers the alignment between the generated and the real responses. WordNet is adopted to calculate the matching relationship among specific sequence matching, synonym, root, interpretation, etc.

NIST [19]. NIST is an improvement of BLEU by summing up each information weighted co-occurrence n-gram segments, then dividing it by the total number of n-gram segments.

There are some other metrics employed in the DGDS, for instance, Qin et al. [73] computed a '#match' score which is the number of non-stop word tokens in the response that are present in the document but not present in the context of the conversation, they also measure the average length of the utterance the CMR model generated. Zhong and Zettlemoyer [136] showed a combined metric ("Comb."), which is the product between the macro-averaged accuracy and the BLEU-4 score on ShARC dataset. Xu et al. [118] adopted Exact Match (EM) in all dialogues. EM requires the answers to have an exact string match with human-annotated answer spans.

6.5 Summarization

We present some evaluation metrics in the current DBD models in Table 13. It should be pointed out that these experimental values can only be used as a reference due to the differences in data processing and experimental environment, more constraints are needed for a fair comparison.

Table 13. The metric employed in the DBD models. * means the test set has two version (Frequent/Rare), we only show the Rare version.

Model	Dataset	F1	BLEU-4	ROUGE-(1/2/L)	Distinct-(1/2)	PPL
QANeT[3]	Holl-E	47.67				
AKGCM[55]	Holl-E		30.84	— — / 29.29 / 34.72		
CaKe[132]	Holl-E		31.16	48.65 / 36.54 / 43.21		
RefNet[62]	Holl-E	48.81	33.65	49.64 / 38.15 / 43.77		
GLKS[82]	Holl-E			50.67 / 39.20 / 45.64		
BiDAF[66]	Holl-E	51.35	39.39	50.73 / 45.01 / 46.95		
SEQS[138]	CMUDoG					10.11
ITEDD[48]	CMUDoG		0.95			15.11
DialogT[100]	CMUDoG		1.28			50.3
TF[29]	T-Chat*	0.20			0.83 / 0.81	43.6
CMR[73]	CbR		1.38		0.052 / 0.283	

7 FUTURE WORK

In this chapter, we discuss some fundamental and technical problems for the future development of the DGDS. Understanding fundamental problems can provide guidence for solving technical problems.

7.1 Fundamental problems

The fundamental problems of the DGDS are listed as below:

- **What is the function of the DGDS?** We divided the function of the DGDS into three levels: the first level is to mine information in document(s), such as the CRC; the second level is to use unstructured document(s) as external knowledge for generating more informative responses, such as the Holl-E; the third level is to take document(s) as discussion objects and express opinions on the contents of document(s), such as the CMUDoG. These three levels can be compared with the process of human learning, using and creating knowledge. The final form of the DGDS should be: the system can refer to the knowledge of the unstructured document(s) or express appropriate views on the document(s) without restraint, and maintain a conversation with users in line with the real human conversation scenario.

- **How to verify the NLU problem?** Do the DGDS models understand the dialogues and the document(s) or they just pick up some unified potential pattern to form a response? Chiang et al. [9] argue that the current CRC models on the QuAC and the CoQA do not well reflect comprehension on conversation content and cross sentence information is not that important. The same question exists in the DBD models. To verify whether the system understands the language and how the system performs the reasoning, we need the DGDS to show a reasonable path of exploring knowledge, rather than simply giving a dialogue reply.
- **How to evaluate the NLG problem?** At present, the NLG tasks usually focus on personalization, diversification, stylization, consistency, etc. We believe that one-to-many problems deserve more attention, especially in the DGDS where entities are restricted to a specific range. We define "one-to-many" as multiple replies involving different knowledge in the document(s) meet the requirements of the dialog context, which contrasts with the normal one-to-one setting. As shown in Table 7, the second utterance of the bot can be each of the four candidates. If candidates A/B/C is the references, D is the generated one, we should train the model to judge its rationality through the utterance history, rather than reduce the probability of its generation because it is different from the other three reference answers. The DGDS datasets [11, 66] nowadays only set multiple references in the test set, failed to train the model with one-to-many properties. The one-to-many problems can better verify the system's understanding and application of natural language. Comparing with the common open domain DS, the diversity of reply in the DGDS can limit to a certain number of entities, which can more realistically achieve one-to-many training and evaluation.
- **Lifelong learning problem.** Most recently, the concept of lifelong learning in the machine learning system has been widely concerned, which requires that the deployed machine learning system continue to improve through interaction with the environment. The idea of lifelong learning is also applicable to the DGDS because it needs to solve the problem of transfer learning not only with different document sources but also with a different task. For example, when the document(s) are news reports, commodity reviews, or novels that come from various data collection sources, the data distribution, syntax structure are different, and the information may also be multimodal. The ShARC task and the CMUDoG task are of some unneglected distinction. Therefore, the DGDS needs to retrain and adjust its strategic components in the deployment process, so that it can automatically learn how to deal with the problem of multi-source heterogeneous data and the distinction between tasks that cannot be completely solved in the training.

7.2 Technical Problems

The technical problems of the DGDS based on the fundamental challenges are listed below:

7.2.1 Memory Ability. Document(s) and historical dialogues should be stored in a long-term memory way. The memory of the document(s) is helpful to judge the utilization of document information in the process of dialogue. The memory of historical dialogue could be used to judge the relationship between historical dialogue information and current dialogue. A multi-turn DS should maintain the memory ability of these two aspects at least, and should not re-model the document(s) and dialogue history every new dialog round. The use of memory ability in the model is not enough.

Liu et al. [55] created an augmented KG with knowledge triples and long texts, which can be regarded as a memory component. They took a factoid knowledge graph (KG) as the backbone, and aligned unstructured sentences of non-factoid knowledge with the factoid KG by linking entities from these sentences to vertices (containing entities) of the KG, augmenting the factoid KG with non-factoid knowledge while retaining its graph structure. Apart from the

ability to remember historical conversations and documents, in order to achieve the final form of DGDS, commonsense knowledge [69, 128, 137] should also exist in memory as a necessary component, so as to form a reasonable dialogue logic and make an appropriate response to the entity which never seen before. Furthermore, we need to store the knowledge representation for lifelong learning.

7.2.2 Reasoning Ability. To verify the NLU, we need the system to be able to show an interpretable reasoning path. It is also the requirement of current AI ethics to be able to reasonably explain the choice of knowledge and reasoning path. Therefore, we believe that the future development trend will depend more on the graph structure which can clearly show the reasoning path and the reinforcement learning method which can explicitly stipulate the reward. There exist several different datasets that require reasoning in multiple steps in literature, for example the google BABI [113], MultiRC [41]and Open-BookQA [63], which are sentence-level reasoning. Welbl et al. [112] and Yang et al. [125] introduced multi-document(s) RC dataset WikiHOP and HotpotQA which need multi-hop reasoning to integrate multi-evidence to locate the target. These studies and the research of the DGDS should benefit each other.

For example, when reasoning in long documents and multiple documents, we can benefit from the current graph-based MRC models. Song et al. [94] used graph convolutional network (GCN) and graph recurrent network (GRN) to better utilize global evidence in WikiHop[112]. Cao et al. [5] directly adopted candidates found in documents as GNN nodes and calculated classification scores over them. Tu et al. [103] introduced the HDE graph containing different types of query aware nodes that represented different granularity levels of information (candidates, documents, and entities) for a multi-choice task. They used GNN based message-passing algorithms to accumulate knowledge on the HDE graph. Lu et al. [60] proposed a KG based QUEST model that computes direct answers to complex questions by dynamically tapping arbitrary text sources and joining sub-results from multiple documents.

7.2.3 One-to-many Porblem. The one-to-one training model using golden reference loses its diversity and generalization, and the current evaluation metrics do not reflect the quality of the model well. In one-to-many training, we need to give appropriate rewards and scores to the generated responses that are not in the reference answers but are correct, which could assist the model to understand natural language and judge the generation ability of the model better. To address the "one-to-many" problem, we need to consider three aspects: (1) Dataset. The multiple reference replies should use different knowledge in the document(s). (2) Training loss function. A good training loss function should have inner relationships with evaluation methods to accomplish better performance. (3) Evaluation. This evaluation metric should meet two requirements: one is to be able to give a reasonable score when the generated reply and multiple reference replies respectively refer to different knowledge sources; the other is to be able to keep consistent with human evaluation. The current methods such as PPL are not suitable for one-to-many settings. Other metrics like BLEU can refer to multiple replies at the same time, but the scenarios in machine translation tasks are usually different from DS.

7.2.4 Model Generalization. In the process of lifelong learning, we need to consider knowledge retention and knowledge transfer. Knowledge retention is defined as the retention of historical experience. Knowledge transfer is defined as the ability to take advantage of the historical experience when dealing with a new type of documents. This requires us to link the modeling process and memory ability, preserve experience knowledge, distinguish good experience from bad one, update the outdated knowledge, and establish new knowledge in the face of unseen tasks, etc. We believe that according to the characteristics of the DGDS tasks and different stages of processing, it is the future trend to build the model on multiple subtasks with multiple levels. Subtasks in different levels learn the commonness among different types of documents. Subtasks in the same level investigate the differences among all types of documents, for

example, merchandise reviews and news report have different text structures, news report can also include multimedia information such as voice, picture, video, etc. There must be different sub-functions to solve different text structures and different data forms.

8 CONCLUSION

The Document Grounded Dialogue System (DGDS) can mine document(s) information and discuss specific document(s) in a real human conversation. We believe that extracting unstructured document(s) information in dialogue is the future trend of the DS because a large amount of human knowledge is contained in these document(s). The research of the DGDS not only possesses a broad application prospect but also facilitates the DS to better understand human knowledge and natural language. This article introduces the DGDS, defines the related concepts, analyzes the current datasets and models, and provides views on future research trends in this field, hoping to be helpful for the community.

REFERENCES

[1] Satoshi Akasaki and Nobuhiro Kaji. 2017. Chat Detection in an Intelligent Assistant: Combining Task-oriented and Non-task-oriented Spoken Dialogue Systems. In *ACL (1)*. Association for Computational Linguistics, 1308–1319.

[2] Satoshi Akasaki and Nobuhiro Kaji. 2019. Conversation Initiation by Diverse News Contents Introduction. In *NAACL-HLT (1)*. Association for Computational Linguistics, 3988–3998.

[3] Siddhartha Arora, Mitesh M. Khapra, and Harish G. Ramaswamy. 2019. On Knowledge distillation from complex networks for response prediction. In *NAACL-HLT (1)*. Association for Computational Linguistics, 3813–3822.

[4] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A Neural Probabilistic Language Model. In *NIPS*. MIT Press, 932–938.

[5] Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question Answering by Reasoning Across Documents with Graph Convolutional Networks. In *NAACL-HLT (1)*. Association for Computational Linguistics, 2306–2317.

[6] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *ACL (1)*. Association for Computational Linguistics, 1870–1879.

[7] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explorations* 19, 2 (2017), 25–35.

[8] Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2019. GraphFlow: Exploiting Conversation Flow with Graph Neural Networks for Conversational Machine Comprehension. *CoRR* abs/1908.00059 (2019). arXiv:1908.00059 http://arxiv.org/abs/1908.00059

[9] Ting-Rui Chiang, Hao-Tong Ye, and Yun-Nung Chen. 2019. An Empirical Study of Content Understanding in Conversational Question Answering. *CoRR* abs/1909.10743 (2019). arXiv:1909.10743 http://arxiv.org/abs/1909.10743

[10] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*. ACL, 1724–1734.

[11] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *EMNLP*. Association for Computational Linguistics, 2174–2184.

[12] Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 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, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 729–738. https://doi.org/10.1145/3357384.3358016

[13] Christopher Clark and Matt Gardner. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *ACL (1)*. Association for Computational Linguistics, 845–855.

[14] Peter Clark. 2015. Elementary School Science and Math Tests as a Driver for AI: Take the Aristo Challenge!. In *AAAI*. AAAI Press, 4019–4021.

[15] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial Paranoia. *Artif. Intell.* 2, 1 (1971), 1–25.

[16] Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on Evaluation Methods for Dialogue Systems. *CoRR* abs/1905.04071 (2019). arXiv:1905.04071 http://arxiv.org/abs/1905.04071

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://www.aclweb.org/anthology/N19-1423/

[18] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational agents. *CoRR* abs/1811.01241 (2018). arXiv:1811.01241 http://arxiv.org/abs/1811.01241

[19] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 138–145.

[20] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.06931

[21] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *NAACL-HLT (1)*. Association for Computational Linguistics, 2368–2378.

[22] Ahmed Elgohary, Chen Zhao, and Jordan L. Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *EMNLP*. Association for Computational Linguistics, 1077–1083.

[23] Alice E. Fischer and Frances S. Grodzinsky. 1993. *The anatomy of programming languages*. Prentice Hall.

[24] Liye Fu, Jonathan P. Chang, and Cristian Danescu-Niculescu-Mizil. 2019. Asking the Right Question: Inferring Advice-Seeking Intentions from Personal Narratives. In *NAACL-HLT (1)*. Association for Computational Linguistics, 528–541.

[25] Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-Again Neural Networks. In *ICML (Proceedings of Machine Learning Research)*, Vol. 80. PMLR, 1602–1611.

[26] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval* 13, 2-3 (2019), 127–298.

[27] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A Knowledge-Grounded Neural Conversation Model. In *AAAI*. AAAI Press, 5110–5117.

[28] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6572

[29] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. *Proc. Interspeech 2019* (2019), 1891–1895.

[30] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *ACL (1)*. The Association for Computer Linguistics.

[31] Yingjie Gu, Xiaolin Gui, and Defu Lin. 2019. TT-Net: Topic Transfer-Based Neural Network for Conversational Reading Comprehension. *IEEE Access* 7 (2019), 116696–116705.

[32] Bin Guo, Hao Wang, Yasan Ding, Shaoyang Hao, Yueqi Sun, and Zhiwen Yu. 2019. c-TextGen: Conditional Text Generation for Harmonious Human-Machine Interaction. *CoRR* abs/1909.03409 (2019). arXiv:1909.03409 http://arxiv.org/abs/1909.03409

[33] Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-Action: Conversational Question Answering Over a Large-Scale Knowledge Base. In *NeurIPS*. 2946–2955.

[34] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *NIPS*. 1693–1701.

[35] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.02301

[36] Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. FlowQA: Grasping Flow in History for Conversational Machine Comprehension. *CoRR* abs/1810.06683 (2018). arXiv:1810.06683 http://arxiv.org/abs/1810.06683

[37] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2019. Challenges in Building Intelligent Open-domain Dialog Systems. *CoRR* abs/1905.05709 (2019). arXiv:1905.05709 http://arxiv.org/abs/1905.05709

[38] Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based Neural Structured Learning for Sequential Question Answering. In *ACL (1)*. Association for Computational Linguistics, 1821–1831.

[39] Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on Conversational Question Answering. *CoRR* abs/1909.10772 (2019). arXiv:1909.10772 http://arxiv.org/abs/1909.10772

[40] Anjuli Kannan and Oriol Vinyals. 2017. Adversarial Evaluation of Dialogue Models. *CoRR* abs/1701.08198 (2017). arXiv:1701.08198 http://arxiv.org/abs/1701.08198

[41] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *NAACL-HLT*. Association for Computational Linguistics, 252–262.

[42] Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie L. Webber. 2017. Edina: Building an Open Domain Socialbot with Self-dialogues. *CoRR* abs/1709.09816 (2017). arXiv:1709.09816 http://arxiv.org/abs/1709.09816

[43] Kalpesh Krishna and Mohit Iyyer. 2019. Generating Question-Answer Hierarchies. In *ACL (1)*. Association for Computational Linguistics, 2321–2334.

[44] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *TACL* 7 (2019), 452–466.

[45] Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *WMT@ACL*. Association for Computational Linguistics, 228–231.

[46] Carolin Lawrence, Bhushan Kotnis, and Mathias Niepert. 2019. Attending to Future Tokens for Bidirectional Sequence Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 1–10. https://doi.org/10.18653/v1/D19-1001

[47] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *HLT-NAACL*. The Association for Computational Linguistics, 110–119.

[48] Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In *ACL (1)*. Association for Computational Linguistics, 12–21.

[49] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[50] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*. The Association for Computational Linguistics, 2122–2132.

[51] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge Diffusion for Neural Dialogue Generation. In *ACL (1)*. Association for Computational Linguistics, 1489–1498.

[52] Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural Machine Reading Comprehension: Methods and Trends. *CoRR* abs/1907.01118 (2019). arXiv:1907.01118 http://arxiv.org/abs/1907.01118

[53] Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic Answer Networks for Machine Reading Comprehension. In *ACL (1)*. Association for Computational Linguistics, 1694–1704.

[54] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[55] Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge Aware Conversation Generation with Reasoning on Augmented Graph. *CoRR* abs/1903.10245 (2019). arXiv:1903.10245 http://arxiv.org/abs/1903.10245

[56] Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhuoran Wang. 2017. A knowledge enhanced generative conversational service agent. In *DSTC6 Workshop*.

[57] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *ACL (1)*. Association for Computational Linguistics, 1116–1126.

[58] Ryan Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, and Joelle Pineau. 2015. Incorporating unstructured textual knowledge sources into neural dialogue systems. In *Neural information processing systems workshop on machine learning for spoken language understanding*.

[59] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL Conference*. The Association for Computer Linguistics, 285–294.

[60] Xiaolu Lu, Soumajit Pramanik, Rishiraj Saha Roy, Abdalghani Abujabal, Yafang Wang, and Gerhard Weikum. 2019. Answering Complex Questions by Joining Multi-Document Evidence with Quasi Knowledge Graphs. In *SIGIR*. ACM, 105–114.

[61] Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging Reading Comprehension on Daily Conversation: Passage Completion on Multiparty Dialog. In *NAACL-HLT*. Association for Computational Linguistics, 2039–2048.

[62] Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. RefNet: A Reference-aware Network for Background Based Conversation. *CoRR* abs/1908.06449 (2019). arXiv:1908.06449 http://arxiv.org/abs/1908.06449

[63] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*. Association for Computational Linguistics, 2381–2391.

[64] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*. ISCA, 1045–1048.

[65] Alexander H. Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 - System Demonstrations*, Lucia Specia, Matt Post, and Michael Paul (Eds.). Association for Computational Linguistics, 79–84. https://www.aclweb.org/anthology/D17-2014/

[66] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *EMNLP*. Association for Computational Linguistics, 2322–2332.

[67] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.), Vol. 1773. CEUR-WS.org. http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[68] Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension. *CoRR* abs/1905.12848 (2019). arXiv:1905.12848 http://arxiv.org/abs/1905.12848

[69] Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine Comprehension Using Commonsense Knowledge. In *SemEval@NAACL-HLT*. Association for Computational Linguistics, 747–757.

[70] Tim Paek. 2001. Empirical Methods for Evaluating Dialog Systems. In *SIGDIAL Workshop*. The Association for Computer Linguistics.

[71] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*. ACL, 311–318.

[72] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*. Association for Computational Linguistics, 2227–2237.

[73] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading. In *ACL (1)*. Association for Computational Linguistics, 5427–5436.

[74] Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019. A Survey on Neural Machine Reading Comprehension. *CoRR* abs/1906.03824 (2019). arXiv:1906.03824 http://arxiv.org/abs/1906.03824

[75] Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with History Answer Embedding for Conversational Question Answering. In *SIGIR*. ACM, 1133–1136.

[76] Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. 2019. Attentive History Selection for Conversational Question Answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1391–1400. https://doi.org/10.1145/3357384.3357905

[77] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf* (2018).

[78] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. (2019).

[79] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP*. The Association for Computational Linguistics, 2383–2392.

[80] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: The Science Behind the Alexa Prize. *CoRR* abs/1801.03604 (2018). arXiv:1801.03604 http://arxiv.org/abs/1801.03604

[81] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *TACL* 7 (2019), 249–266.

[82] Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation. *CoRR* abs/1908.09528 (2019). arXiv:1908.09528 http://arxiv.org/abs/1908.09528

[83] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of Natural Language Rules in Conversational Machine Reading. In *EMNLP*. Association for Computational Linguistics, 2087–2097.

[84] Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph. In *AAAI*. AAAI Press, 705–713.

[85] Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-Evaluating ADEM: A Deeper Look at Scoring Dialogue Responses. In *AAAI*. AAAI Press, 6220–6227.

[86] Sashank Santhanam and Samira Shaikh. 2019. A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions. *CoRR* abs/1906.00500 (2019). arXiv:1906.00500 http://arxiv.org/abs/1906.00500

[87] Ruhi Sarikaya. 2017. The Technology Behind Personal Digital Assistants: An overview of the system architecture and key components. *IEEE Signal Process. Mag.* 34, 1 (2017), 67–81. https://doi.org/10.1109/MSP.2016.2617341

[88] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL (1)*. Association for Computational Linguistics, 1073–1083.

[89] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional Attention Flow for Machine Comprehension. *CoRR* abs/1611.01603 (2016). arXiv:1611.01603 http://arxiv.org/abs/1611.01603

[90] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A Survey of Available Corpora For Building Data-Driven Dialogue Systems: The Journal Version. *D&D* 9, 1 (2018), 1–49.

[91] Abhishek Sharma, Danish Contractor, Harshit Kumar, and Sachindra Joshi. 2019. Neural Conversational QA: Learning to Reason v.s. Exploiting Patterns. *CoRR* abs/1909.03759 (2019). arXiv:1909.03759 http://arxiv.org/abs/1909.03759

[92] Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. ReasoNet: Learning to Stop Reading in Machine Comprehension. In *KDD*. ACM, 1047–1055.

[93] Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of IT & EE* 19, 1 (2018), 10–26.

[94] Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018. Exploring Graph-structured Passage Representation for Multi-hop Reading Comprehension with Graph Neural Networks. *CoRR* abs/1809.02040 (2018). arXiv:1809.02040 http://arxiv.org/abs/1809.02040

[95] Lixin Su, Jiafeng Guo, Yixing Fan, Yanyan Lan, Ruqing Zhang, and Xueqi Cheng. 2019. An Adaptive Framework for Conversational Question Answering. In *AAAI*. AAAI Press, 10041–10042.

[96] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A Challenge Dataset and Models for Dialogue-Based Reading Comprehension. *TACL* 7 (2019), 217–231.

[97] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*. 3104–3112.

[98] Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *NAACL-HLT*. Association for Computational Linguistics, 641–651.

[99] Alon Talmor, Mor Geva, and Jonathan Berant. 2017. Evaluating Semantic Parsing against a Simple Web-based Question Answering Model. In **SEM*. Association for Computational Linguistics, 161–167.

[100] Xiangru Tang and Po Hu. 2019. Knowledge-Aware Self-Attention Networks for Document Grounded Dialogue Generation. In *KSEM (2) (Lecture Notes in Computer Science)*, Vol. 11776. Springer, 400–411.

[101] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *AAAI*. AAAI Press, 722–729.

[102] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Rep4NLP@ACL*. Association for Computational Linguistics, 191–200.

[103] Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop Reading Comprehension across Multiple Documents by Reasoning over Heterogeneous Graphs. In *ACL (1)*. Association for Computational Linguistics, 2704–2713.

[104] Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the Turing Test*. Springer, 23–65.

[105] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[106] Pavlos Vougiouklis, Jonathon S. Hare, and Elena Simperl. 2016. A Neural Network Approach for Knowledge-Driven Response Generation. In *COLING*. ACL, 3370–3380.

[107] Marilyn A. Walker, John S. Aberdeen, Julie E. Boland, Elizabeth Owen Bratt, John S. Garofolo, Lynette Hirschman, Audrey N. Le, Sungbok Lee, Shrikanth Narayanan, Kishore Papineni, Bryan L. Pellom, Joseph Polifroni, Alexandros Potamianos, P. Prabhu, Alexander I. Rudnicky, Gregory A. Sanders, Stephanie Seneff, David Stallard, and Steve Whittaker. 2001. DARPA communicator dialog travel planning systems: the june 2000 data collection. In *INTERSPEECH*. ISCA, 1371–1374.

[108] Marilyn A. Walker, Rebecca J. Passonneau, and Julie E. Boland. 2001. Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems. In *ACL*. Morgan Kaufmann Publishers, 515–522.

[109] Marilyn A. Walker, Alexander I. Rudnicky, John S. Aberdeen, Elizabeth Owen Bratt, John S. Garofolo, Helen Wright Hastie, Audrey N. Le, Bryan L. Pellom, Alexandros Potamianos, Rebecca J. Passonneau, Rashmi Prasad, Salim Roukos, Gregory A. Sanders, Stephanie Seneff, and David Stallard. 2002. DARPA communicator evaluation: progress from 2000 to 2001. In *INTERSPEECH*. ISCA.

[110] Joseph Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[111] Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *NUT@EMNLP*. Association for Computational Linguistics, 94–106.

[112] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *TACL* 6 (2018), 287–302.

[113] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1502.05698

[114] Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory Networks. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1410.3916

[115] Jason Weston, Emily Dinan, and Alexander H. Miller. 2018. Retrieve and Refine: Improved Sequence Generation Models For Dialogue. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail Burtsev (Eds.). Association for Computational Linguistics, 87–92. https://www.aclweb.org/anthology/W18-5713/

[116] Xianchao Wu, Ander Martinez, and Momo Klyen. 2018. Dialog Generation Using Multi-Turn Reasoning Neural Networks. In *NAACL-HLT*. Association for Computational Linguistics, 2049–2059.

[117] Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation Networks: Sequence Generation Beyond One-Pass Decoding. In *NIPS*. 1784–1794.

[118] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Review Conversational Reading Comprehension. *CoRR* abs/1902.00821 (2019). arXiv:1902.00821 http://arxiv.org/abs/1902.00821

[119] Zhen Xu, Nan Jiang, Bingquan Liu, Wenge Rong, Bowen Wu, Baoxun Wang, Zhuoran Wang, and Xiaolong Wang. 2018. LSDSCC: a Large Scale Domain-Specific Conversational Corpus for Response Generation with Diversity Oriented Evaluation Metrics. In *NAACL-HLT*. Association for Computational Linguistics, 2070–2080.

[120] Rui Yan. 2018. "Chitty-Chitty-Chat Bot": Deep Learning for Conversational AI. In *IJCAI*. ijcai.org, 5520–5526.

[121] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. In *AAAI*. AAAI Press, 4618–4626.

[122] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A Hybrid Retrieval-Generation Neural Conversation Model. In *CIKM*. ACM, 1341–1350.

[123] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. In *SIGIR*. ACM, 245–254.

[124] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *CoRR* abs/1906.08237 (2019). arXiv:1906.08237 http://arxiv.org/abs/1906.08237

[125] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP*. Association for Computational Linguistics, 2369–2380.

[126] Mark Yatskar. 2019. A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. In *NAACL-HLT (1)*. Association for Computational Linguistics, 2318–2323.

[127] Yi Ting Yeh and Yun-Nung Chen. 2019. FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension. *CoRR* abs/1908.05117 (2019). arXiv:1908.05117 http://arxiv.org/abs/1908.05117

[128] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge. In *AAAI*. AAAI Press, 4970–4977.

[129] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *ACL (1)*. Association for Computational Linguistics, 2204–2213.

[130] Xin Zhang, An Yang, Sujian Li, and Yizhong Wang. 2019. Machine Reading Comprehension: a Literature Review. *CoRR* abs/1907.01686 (2019). arXiv:1907.01686 http://arxiv.org/abs/1907.01686

[131] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. In *NeurIPS*. 1815–1825.

[132] Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2019. Improving Background Based Conversation with Context-aware Knowledge Pre-selection. *CoRR* abs/1906.06685 (2019). arXiv:1906.06685 http://arxiv.org/abs/1906.06685

[133] Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. In *COLING*. Association for Computational Linguistics, 3740–3752.

[134] Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskénazi. 2017. Generative Encoder-Decoder Models for Task-Oriented Spoken Dialog Systems with Chatting Capability. In *SIGDIAL Conference*. Association for Computational Linguistics, 27–36.

[135] Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. 2019. A Document-grounded Matching Network for Response Selection in Retrieval-based Chatbots. In *IJCAI*. ijcai.org, 5443–5449.

[136] Victor Zhong and Luke Zettlemoyer. 2019. E3: Entailment-driven Extracting and Editing for Conversational Machine Reading. In *ACL (1)*. Association for Computational Linguistics, 2310–2320.

[137] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*. ijcai.org, 4623–4629.

[138] Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A Dataset for Document Grounded Conversations. In *EMNLP*. Association for Computational Linguistics, 708–713.

[139] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *CoRR* abs/1812.08989 (2018). arXiv:1812.08989 http://arxiv.org/abs/1812.08989

[140] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In *ACL (1)*. Association for Computational Linguistics, 1118–1127.

[141] Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering. *CoRR* abs/1812.03593 (2018). arXiv:1812.03593 http://arxiv.org/abs/1812.03593