International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:14, No:12, 2020

A Survey of Response Generation of Dialogue Systems

Yifan Fan, Xudong Luo, Pingping Lin

Abstract—An essential task in the field of artificial intelligence is to allow computers to interact with people through natural language. Therefore, researches such as virtual assistants and dialogue systems have received widespread attention from industry and academia. The response generation plays a crucial role in dialogue systems, so to push forward the research on this topic, this paper surveys various methods for response generation. We sort out these methods into three categories. First one includes finite state machine methods, framework methods, and instance methods. The second contains full-text indexing methods, ontology methods, vast knowledge base method, and some other methods. The third covers retrieval methods and generative methods. We also discuss some hybrid methods based knowledge and deep learning. We compare their disadvantages and advantages and point out in which ways these studies can be improved further. Our discussion covers some studies published in leading conferences such as IJCAI and AAAI in recent years.

Keywords—Retrieval, generative, deep learning, response generation, knowledge.

I. Introduction

WITH the continuously rapid development of computer science, the way of human-computer interaction has undergone tremendous changes. From the traditional keyboard and mouse to the touch screen, voice, and other methods, the means of human-computer interaction become more and more convenient for humans [55]. Interaction via language is one of the necessary skills of human beings, so it is also the most straightforward way humans communicate with computers. If humans can communicate with computer systems through natural language, they can conveniently access various information that computer systems provide. A dialogue system not only bridges humans to a computer but is also one of the most typical natural language processing applications.

Dialogue systems have a promising prospect, and various domains demand such systems very much. In the past decade, dialogue systems have been widely applied in entertainment, navigation, and communication in different forms, such as:

- *Personal assistant systems* (*e.g.*, Google Now [9], Siri [2], and Baidu's smart assistant secretary). Their dialogue functions can effectively save users considerable time and effort when they access various services. For example, through speech, users can instruct a personal assistant system to complete the functions that users need to click multiple times (*e.g.*, check the weather, search movie, and book services). And in the elderly care industry, the accompanying robots with chat functions can chat with the elderly and help them to access online services easily.
- *Voice control systems*. By such systems, human users can use speech to interact with home appliances at home, realise natural language interaction with cars, and help users to complete various functions. For example, in the current smart home [46], a human user can voice-activate various operations of home appliances, such as users can ask TV to play movies or news, and operate the basic functions of electronic devices. Apple's CarPlay system [13] is an in-vehicle system based on iOS. The car with the system can realise voice calls, SMS sending and receiving, and electronic equipment operation through natural language interaction function at any time.
- *Customer service robots* [35]. The examples of them include Taobao's Ali Xiaomi and Jingdong's intelligent customer service robots. These customer service robots can help users solve their problems in shopping and respond to frequently used functional and repetitive questions through automated responses. Customer service robots based on human-computer dialogue systems can effectively reduce the pressure of customer service personnel. They can significantly improve the efficiency of the customer service, save the cost of human resources, and guarantee more effective services in shopping platforms.

A dialogue system consists of three parts: understanding what humans say in natural language, managing dialogue, and generating responses in natural language. The main task of dialogue management is to detect the state of the dialogue and take proper dialogue strategies to ensure that a conversation can be carried out efficiently towards the preset objectives. A proper dialogue management function can accurately grasp the progress of the dialogue, and continuously optimise the quality of communication in the process of multiple rounds of conversation with the user, thereby improving user satisfaction and bringing higher profits to the merchant. The natural language response generation of a dialogue system is also an essential component of a dialogue system. A good generation model can response to the user with accurate information by generating a sentence suitable for the new scene with a small amount of training corpus [32].

In the early days of developing dialogue systems, the response generation was relatively simple. In recent years,

This work was supported by the National Natural Science Foundation of China (No. 61762016) and Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (No. 19-A-01-01).

Yifan Fan, Xudong Luo*, and Pingping Lin are with Guangxi Key Lab of Multi-Source Information Mining & Security, College of Computer Science and Information Engineering, Guangxi Normal University, Guilin 541004, China (*corresponding author, e-mail: luoxd@mailbox.gxnu.edu.cn).

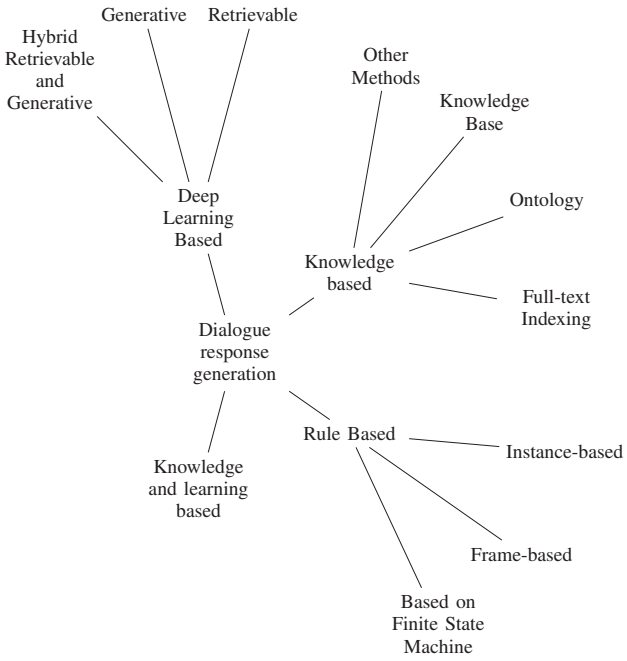

Fig. 1 The Classification of response generation methods

the success of Go Man-Machine Warfare has attracted the attention of industry scholars, and the design of the module of response generation has taken a big step forward. The response generation methods are sorted into three categories: retrieval formula, generative formula, and the method of combining retrieval and generation, as summarised in Fig. 1.

In this paper, we focus on surveying the methods of response generation of dialogue systems. In existing survey papers on dialogue systems, the researchers just try to cover all aspects of dialogue systems and thus brief some methods of response. Chen et al. [7] survey dialogue systems from the task-based and non-task-based viewpoint, but However, they just brief several response generation methods based on deep learning methods. Chen et al. [6] survey the open-domain dialogue system as the research object, it mainly conducts a comprehensive analysis and summary of the implementation methods of deep learning for a dialogue system. However, the research contents involved were published before 2019, and the latest research results have not been involved. Instead, in this paper, we focus us on the updated methods of response generation in task-based dialogue system and open-domain dialogue system. More specifically, we survey the studies on this topic which were published in the past three years, especially those in the top journal and the top conference. In particular, we compare the characteristics of different response generations and standard model implementation methods, and we also point out future research directions.

The rest of this paper organised as follows. Section II discusses threes kinds of rule-based method for response generation. Sections III focuses on threes kinds of knowledge-based methods. Section IV examines some deep learning based methods. Section V briefs some hybrid methods based on knowledge and deep learning. Finally, Section VI concludes this paper with future work.

II. Rule Based Methods

This section will discuss three kinds of response generation methods based on rules.

A. Basic Principle

Generally speaking, such a method is to set some dialogue scenes manually and then write a corresponding dialogue template for each scene. The domain model and these dialogue templates can guide users to complete specific tasks, so it has a high success rate.

In a dialogue system ALICE (Artificial Linguistic Internet Computer Entity) [56], a designer, use AIML language to describe the dialogue templates. The AIML language store data in the XML language. AIML defines the data format and corresponding operations of the object, and also allows the user to extend the system of ALICE functionality according to the designer's definition of new tags, which is very scalable. Also, if using a template matching algorithm to find the answer to a question, we need to design the possible ground problems artificially, and their responses. The design difficulties of the specific domain and the open-domain are different.

In a question and answer (Q&A) systems, answers for FAQ (Frequently Asked Question) [51] need to be manually acquired in a specific field and stored as pairs of questions and answers in a relational database. In such a dialogue system, as long as a user's question can match the question template existing in the system, the stored corresponding response is directly displayed to the user. In the communication of dialogue, this form requires the system to add some specific parameters according to the particular conversation. The advantage is that the answer is accurate in a specific field. The disadvantage is that it requires too much manual labelling and template writing, which leads to the problems of portability and expandability, poor performance, lack of flexibility.

B. Finite State Machine Based Methods

The method based on Finite State Machine (FSM) is the earliest widely used response generation method. It is a method of rule-based response generation. It is suitable for some dialogue systems with simple and precise requirements. The FSM model has a wide range of application scenarios for machine tool control modules in the engineering field [20]. A finite number of states can present all possible conversation processes, and a dialogue process regards it as the state transition process of the automaton. The system needs to define all possible conversation states in advance and formulate a corresponding state transition mechanism, which is a finite conversation state network. According to the recognition of the user's voice intention, Shen and Inkpen [43] use a finite state machine to process the conversion of the corresponding sub-module, so that the system and its users can maintain an orderly session. In the process of dialogue, each interactive node is predefined, each state corresponds to a prompt message

of the system, and it expects to obtain feedback from its users. According to the information input of the user, the system jumps to a new one. However, such FSM methods are only applicable to a simple system and are not suitable for a complex interactive system.

The designer of an FSM model is required to list all possible dialogue states and possible responses to its users at the time of design (*i.e.*, the conditions for the transfer of all states must be known). In general, a finite state machine model can use a quintuple:

$$< M = Q, q_0, \sum, \delta, F >$$

where Q is a finite state set; $q_0 \in Q$ represents the initial state; $\sum$ is a finite event input set; $\delta : Q \times \sum \rightarrow Q$ is a state transfer function (in particular $\delta(q, a) = q'$ means that if the input event is a at state q , the machine will enter state q'); and F is the termination state set and $F \subseteq Q$. So we can see that an FSM model has a clear structure (so it is easily implemented) and can well control the flow of a user's interaction with the system. Thus, a lot of theoretical studied has used it. For example, in the framework of the human-machine interface task manager, Lee et al. [21] use an FSM to manage effectively the generation services provided by each robot; Abe et al. [1] propose a new method of response generation for intelligent retrieval systems, constructing two independent finite state machines to simulate users and systems.

Some researchers use FSMs for generating a reasonable response in their dialogue systems. For example, Raux and Eskenazi [37] introduce a kind of FSN model and choose a dialogue strategy by setting the cost matrix of the action and minimising the expected cost of the action. In the online shopping scenario, if the status of the dialogue is *Recommended*, the action with respect to *Recommended* is triggered, and the system retrieves the product from the product database. If the state is *Compare*, the system compares the product in the product database with the target product. The system will give corresponding responses according to different trigger actions [63].

FSM based models can perform excellently in different fields. For a task-based dialogue system, it is a vertical service structure. Designers can predict all possible situations according to specific tasks. However, when the task area to which it belongs changes, the response generation part may need to redesign. In the case of an open domain dialogue system, the domain of the problem design by the user is complicated. Actually, it is a big challenge to ensure that in the state diagram, there are no errors; and it consumes many workforces to implement it. It is also tricky for a response generation method based on FSM to deal with unforeseen circumstances. If the user's response is entirely beyond the scope of the response generation module, the dialogue will not be able to proceed normally. It is not applicable in open-domain dialogue systems.

C. Framework Based Method

Frame-based response generation is a kind of sheet-like tasks, also known as channel filling (Slot-Filling) method or a sheet. It is one new response generation method often used. It completes the information set that needs to be determined on the table through communication with a human user until the unknown information on the table is available. Most of the first information query is the method of filling in the form. The process of filling the form of a topic task can be represented by the form of a table. The dialogue is a form filling process according to the established dialogue rules (*i.e.*, elicit the information through constant interaction with a user). The padding of the form determines the corresponding entry in the table, and the dialogue policy for the next session. The human-computer interaction level of the dialogue system realised by the method of filling out the form is not high. In this way, the dialogue is relatively mechanical, unnatural, but the implementation complexity is low.

Goddeau et al. [11] propose a dialogue planning algorithm based on the concept of a spreadsheet, using the form filling method to establish a car dialogue advertising system. The designed spreadsheet contains the parameters of the car manufacturer, model, price, mileage, and colour. These parameters include the required necessary fields, and different values will be a corresponding groove prompts to guide the user to give the relevant valid response. Wu et al. [58] propose a goal-driven dialogue system probability framework. By defining a new random dynamic to describe the dialogue state objectives at different stages of a dialogue process, the system can be more effective in achieving the user's goals. Oh et al. [34] use an incremental learning method based on the training corpus to build a response generation model. Its internal data structures in the conversation model base on a semantic representation framework. The results of speech understanding are groove set values. The content-type slot values contained in the program has starting time and actors. They use incremental learning methods based on training corpora to construct dialogue strategies to form a reasonable response generation semantic framework.

Compared with FSM based methods, the framework-based response generation methods do not need to consider the dialogue order. A conversation can ignore the information irrelevant to the content. For some critical information, the information can be confirmed with a user to ensure accurate identification. When the user gives the information in advance for specific questions, the system must be able to discover and obtain the information. Therefore, the framework-based response generation can widely be applied in business dialogue systems. However, it may lead to a very mechanical dialogue process, and the interaction between the user and the system may not be very natural. When dealing with multi-theme dialogue tasks, such a method is not ideal.

After we understand the advantages and disadvantages of the two kinds of methods, we may think of integrating the two to design a hybrid method. One possible way for integrating the two is to use an FSM based model to control the entire conversion process, then define several states that need to interact with the user, and implement the conditions of the system state transition by using the slot result. On the line of this idea, Hurtado et al. [14] use a stochastic FSM to model dialogue response generation by uncompressing all the information provided by the user into

a state throughout the previous history of the conversation. Their method uses a training corpus to build up the response generation model automatically. In order to better build up the response generation model, they define some specific aspects of tasks, for example, the semantics of a task and a set of dialogue actions between the system and users. In general, the research of early dialogue systems tends to integrate FSM based methods and framework based methods.

Improving the accuracy of the speech recognition and speech understanding modules of a dialogue system can help the response generation model based on the framework to make better decisions. Shi and Huang [45] propose a deep sequential model for discourse analysis in multi-party dialogues. This model sequentially scans Basic Language Units (BLUs) in the dialogue. For each BLU, the model determines that the current BLU should link to which BLU and what is the corresponding relationship type, and then use the predicted link and relationship type to construct the discourse structure through the structured encoder gradually. In the use of predicted links and relationship classification, the model uses not only the local information of the relevant BLU but also the global information that encodes the BLU sequence and the established utterance structure in the current step. As future work, this method can be enhanced and applied to other Natural Language Process (NLP) tasks that improve multi-party conversations, such as intelligent customer service, chat robots, and other applications.

D. Instance-Based Methods

An instance-based method for response generation is simple but useful for many dialogue systems. Mainly, it bases the idea of pattern matching. It matches the problem in the database according to the user's input and then gives the corresponding response strategy. The response generation method based on pattern matching does not have a context so that it can be applied in open-domain dialogue systems. Furthermore, its requirements are simple so that it is easy to implement. For a task-based dialogue system, in the case that the requirement for the quality of a dialogue is not very high, it is fine to use this method.

Researchers have proposed several instance-based methods for response generation. Lee et al. [22] propose an instance-based response generation method mainly for specific areas with excellent domain transplant characteristics. According to the user input language, they offer different methods for semantic understanding, keyword characteristics, and the domain classification. Lee et al. [23] propose a multi-domain response generation approach. It can manage task-based dialogue and entertainment chat in areas such as car navigation, weather information, and TV program guidance. They use real and simulated users to evaluate their dialogue system. Kim et al. [17] propose an instance-based response generation model. In the system, the dialogue state hypothesis and the confirmation strategy determine whether or not use the confirmation mechanism. If the user's input appears frequently, the scope of the confirmed belief needs to narrow down. The confirmation mechanism provides a solution for the natural and efficient progression of the conversation. Noh et al. [33] propose a hybrid of the response generation method, using a statistical sorting algorithm based on instances of dialogue response. By analysing a causal relationship between the behaviour of a user's voice and intention, the dialogue system can predict the intent that the user may want to express. The statistical sorting algorithm uses three features to realise statistics: the similarity of ranking statements, the order of causality between language behaviours, and the state of filling of entities.

In an open-domain, users may raise a wide range of questions. If an open-domain system use pattern matching for response generation, it may need a large number of question and answer templates, which elicitation can consume a lot of workforces and resources. However, some answers may still be unnatural, so users are likely reluctant to participate in the conversation. As a result, this response generation method has significant limitations. It cannot adapt to changing sentences and has a poor ability to analyse complex sentence structures. Thus, it is just suitable for some dialogue systems where the accuracy of responses is not critical. Compared with statistical methods, the instance-based method does not consider the strategy of using system optimisation. It relies mainly on heuristic similarity metrics in the corpus to find a correct instance.

III. Knowledge Based Method

In this section, we are going to discuss the knowledge based methods for response generation.

A. Full-Text Indexing

In the current field of Q&A systems, researchers widely use full-text indexing technology for response generation [59]. The technology uses knowledge base retrieval, so it is somewhat similar to the design of a search engine. A search engine gets keyword input in the search box, and then its interface returns the pages that, it thinks, are most relevant to the keyword(s). The response generation method based on the knowledge base retrieval technology needs a knowledge base prepared in advance. The knowledge base should contain precious dialogue materials, and the questions users asked previously. The system should be able to analyse the questions raised by the users by using the technology of NLP keyword extraction, inverted index, document sorting, and other methods for fuzzy matching with the knowledge base to find the appropriate response content. The advantage of this method is that the quality of the answer is very high, and the expression is natural. Its disadvantage is that it is hard to find rich data and proper knowledge. Also, because of the large amount of data, it is not very easy to manage a dialogue continuously.

B. Ontology

The ontology-based knowledge base is also widely used for response generation. Ontology conceptualises all real-world entity relationships, concepts, and attributes, and represents the relationship between entities in the form of structured data

plus relevant domain axioms. So such a system can perform intelligent reasoning. In a dialogue system, if we integrate the concept of ontology system with the artificially established rule system, it can provide the source of knowledge and inference for responses to the questions asked by users.

In most cases, there are semantically similar query keywords, and different methods used in grammar. For example, regarding the standard terminology in the service industry, buy and purchase (synonym) have the same meaning. Words, which express different meanings, such as order, rely solely on the relationship between keywords. Ignoring its semantics, sometimes, results in that the dialogue is unsatisfactory. To provide users with abundant and caring services, Broens et al. [5] use ontology-based retrieval to classify users' services according to their attributes, instead of using simple keyword retrieval methods, and based on ontology to specify context entities. The relationships allow these entities to be shared and provide better service.

C. Vast Knowledge Base Method

Asakiewicz et al. [3] develop a system (called DeepQA), which can answer different types of customers, such as potential applicants, newly admitted students, faculty, students, and business professionals. The DeepQA system is different from the transaction systems of the programming era as well as Web-based query systems. It processes unstructured data in the form of a vast knowledge base. The source of the knowledge base can be specially developed text, publications, or websites. Resources that support natural language queries can become smarter over time. DeepQA is a knowledge based Q&A system that uses rule-based in-depth grammar analysis and statistical classification methods to determine whether or not to decompose a problem and how to solve it to get responses quickly for each branch after decomposition. A complete set of alternative response generation processes can give a definitive response to a user based on a sub-question through search and quantitative evaluation. The process, policy, or course questions provide an accurate response, and over time, the system grows in coverage and usability.

This kind of structure obviously cannot effectively keep up with the growth of the source knowledge, and cannot effectively interact with users, and cannot provide decisions for users with the support of a large amount of unstructured content, either. Moreover, its dialogue fluency is weak and very likely; its response does not correspond to the problem. However, the advantage of this method is that it is easy to expand its knowledge base, and its response has no grammatical errors.

D. Other MMethods

Wen et al. [57] and Bordes, Boureau, and Weston [4] view the problem of response generation as a mapping from dialogue history to a response. However, due to the lack of extensive and high-quality data of dialogue history, it does not scale well.

Inspired by key-value storage networks, Eric and Manning [10] establish an attention-based key-value retrieval mechanism based on each entry in the knowledge base, which enhanced the existing cyclic network architecture.

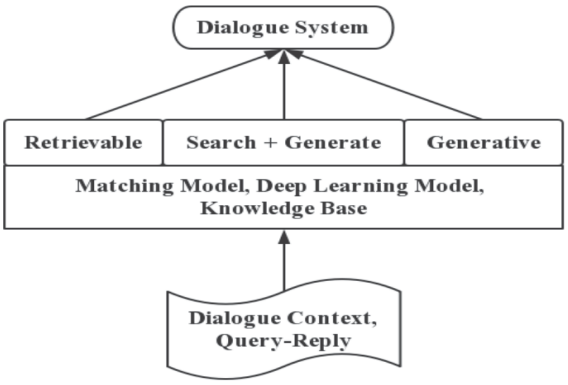

Fig. 2 Response generation based on Deep Learning

The soft posterior distribution and symbol query replacement caused by Dhingra et al. [8] in the knowledge base indicate the entities that the user is interested in, combining the soft retrieval process with reinforcement learning methods.

Serban et al. [39] integrate natural language generation models and retrieval models, including template-based models, bag-of-words models, sequence-to-sequence neural networks, and hidden variable neural network models. It uses reinforcement learning to crowdsourced data and interacts with users. It is better to select the appropriate response from its collection of models.

IV. Deep Learning Based Method

This section will discuss the deep learning based method for the response generation of dialogue systems.

A. Overview

With the success of deep learning research, the model design of the dialogue system has a new research direction. Notably, the research of end-to-end dialogue model in the generative form in deep learning. Conventional deep learning techniques include neural network models (NN), self-encoders, convolutional neural networks (CNN), recurrent neural networks (RNN), sequence-to-sequence (seq2seq), attention mechanisms, memory networks, reinforcement learning (RL), and generational adversarial networks (GAN). Table I summarises the research on the generation module of the dialogue system based on deep learning.

The response generation methods are sorted into three categories: retrieval formula, generative formula, and the method of combining retrieval and generation. Table II compares the threes kinds of response generation methods in several aspects. The schematic diagram of the dialogue frame based on these three methods shows in Fig. 2.

The focus of the current research of deep learning based methods for response generation is on how to use data-driven and related algorithms to generate natural conversations efficiently.

TABLE I
CLASSIFICATION OF RESPONSE GENERATION FRAMEWORK BASED ON DEEP LEARNING

Method	Classification	Features	The role of models	Commonly used models
Retrievable	Representation-centric framework	Capturing feature information in the semantic space	Build query-response text semantic representation model	CNN, RNN, attention mechanism, self-encoder
	Fusion-centric framework	Capture semantic fusion information between query and response	Modeling the query-response semantic fusion process	DNN, CNN, RNN, attention mechanism, GRU
	representation + fusion framework	The combination of query-response feature and semantic fusion	Construction and combination of sub-models	CNN Less research
	Reordering-centric framework	Sort candidate responses to queries	Model text in interactions	RNN, CNN, deep reordering
Generative	seq2seq model framework	Generate a response for a given query	Learn the corresponding mode between user input and response	seq2seq, attention mechanism, intention network, multi-channel coding
	Neural Language Model Framework	Given a word to predict the next word	Each generated result refers to a different input	RNN, attention mechanism
	RL framework	Interact with the environment and guide and improve the dialogue system	Integrate reward strategies and establish long-term impact mechanisms	RL, GAN
	Hierarchical seq2seq model framework	Extract valuable information with historical dialogue text	Extract valuable information with historical dialogue text	Hierarchical seq2seq, memory network

B. Retrieval Method

A retrieval based response generation method does not create a new response, but instead stores the response data in advance, and then finds a reasonable response from the database according to specific rules or relying on certain knowledge.

Such a method can be implemented with a deep learning model. It consists of a query-response matching model, a semantic representation model, and a semantic fusion model. The semantic representation model is to obtain the semantic vector of query and response, and the semantic fusion model is to model the process of query and response semantic vector fusion. A deep learning method can be used to enhance the ability of semantic representation and semantic fusion calculation.

TABLE II
COMPARISON OF THREE DIALOGUE METHODS

Method	Method to realise	Features	Advantages	Disadvantages
Search	Matching model based on deep learning	Simple idea and modular design	Simple idea and modular design	Influence between dialogue modules, single response
Generative	deep learning, the combination of various models	End-to-end model design	With enough corpus training, generate multiple responses without searching	Obtaining corpus, training is difficult, and the quality of response is uncontrollable
Search and generation	Retrieve model candidate answers as the basis for generating model training	There are different ways to optimise an input	High-quality responses, few grammatical errors, and diverse responses	Retrieval model and generative model fusion method

Some studies of this kind are as follows. Shen et al. [44] use the query and candidate response word vectors as input to a Convolutional Neural Network (CNN). Through convolution and pooling operations, it can obtain a fixed-length semantic representation vector. Finally, the cosine similarity function is used to find a response which is the most similar to that of the query. Wan et al. [52] use Bi-LSTM to calculate the sentence representation vector and uses the Bi-LSTM hidden vector sequence to represent the semantics of the sentence from different angles. In the process of semantic fusion, it match semantic representations in different contexts, so that it can calculate the matching score from multiple angles. Yin et al. [66] use CNN to express question and response pairs semantically and uses the attention mechanism to obtain the representation of another sentence based on one sentence representation, thereby expanding the representation information of the sentence. Moreover, Lu and Li [29] use a deep neural network to implement a deep matching model focusing on fusion. It uses a corpus to train topic models at different levels, detects the topics expressed by the text to be matched, and finally, according to these topics find proper responses to given queries. The connection between them, construct a semantic fusion matrix, and finally, use the deep neural network to calculate the matching score to get the best matching response. Zhang et al. [68] use deep learning models to achieve semantic representation and semantic fusion. Both use a self-attention mechanism and a Gate Collection Unit (GRU). They are using word-based and sentence-based granularity to obtain the semantic representation of utterance and two matching matrices. Pooling and convolution methods can obtain matching feature vectors for utterances and queries,

TABLE III
COMBINING HISTORICAL DIALOGUE INFORMATION AND QUERY METHODS

Combination method	Manifestations	Advantages	Disadvantages
Direct stitching	Word sequence	Simple to implement	Ignore the correlation between the two and introduce noise
Consolidation related information	Utterance sequence	Reduce the impact of queries on responses	Not suitable for large amounts of information
Expand queries based on historical conversation	Extended query	Expand limited query combinations to improve performance	High complexity and wide range of exhaustion
Conversation history and query	Semantic representation	Consider the contextual relevance to reduce the impact of noise	High complexity and difficult combination

and input into GRUs in sequence can calculate candidate matching response scores.

Zhou et al. [71] only use the attention mechanism to implement the retrieval and matching model of the dialogue system. In the process of semantic representation, the self-attention mechanism is used to obtain the query-response representation matrix. Semantic fusion obtains two matching matrices by constructing self-attention matching and cross-attention matching to capture the matching features of context and query. Finally, the combination of matching matrices which are pooled to obtain a suitable matching score.

Kim et al. [18] propose a model that contains multiple layers of RNN+Attention structure to implement the matching module of the dialogue system. Due to the spliced connection method, the dimension of the vector increase, but the gradient disappears. It also uses an autoencoder algorithm to reduce the input dimension of each layer of the network. Therefore, it is not only able to eliminate the disappearance of the gradient but also normalise and reduce the complexity of the model.

Song et al. [49] propose a reordering model for response generation. It first retrieves a series of response candidates for a query. Then it reorders the response candidates based on the user and system interaction information. The final response is related to the query and context. In the system, a deep learning model such as CNN and Recurrent Neural Network (RNN) is used to model sentences. They use deep learning only to screen out response candidates related to the query, and there is not much work on the impact of the conversation context on responses.

Deep learning semantic representation models in retrieval dialogues use recurrent neural networks, convolutional neural networks, and attention mechanisms. Unlike a single round of dialogue, multiple rounds of retrieval dialogues need to model the combination of actual dialogue utterances and queries. Table III summarises several combination methods.

Dialogue systems based on retrieval methods generate responses by using the words that appear in the real dialogue, so the quality of the response sentences is high, and their grammatical errors are few. A large-scale corpus is a key to the success of a retrieval dialogue model. Even if the retrieval storage system is extensive, but the query text does not give an appropriate response, the dialogue system cannot generate a new response. The model based on semantic representation is relatively independent of the process of response generation. Although some features may be lost, the effect of matching simple problems is good. The fusion-based dialogue model extracts valuable information in the context and the query response. For complex problems, the final matching effect is good, but the algorithm is complicated, and the calculation amount is significant.

Regarding the use of deep learning for the response generation of a dialogue system, the key to success is to build up a suitable matching model. RNN can be used to calculate the matching degree according to the context utterance order. CNN can directly extract features according to the matching matrix. Recent research shows that the model that introduces the self-attention mechanism has an excellent matching performance for retrieval dialogue systems [60]. Regarding the selection of retrieval methods, the step-by-step training that combines the advantages of different models can improve the performance of such a dialogue system, whether it is in the efficiency of search matching or the quality of matching in answering user questions.

C. Generative Method

In a generative method based on the framework of deep learning, through the training of query-response corpus, in the prediction stage, a response can be generated according to the semantic vector input by the user.

Many researchers now pursue the goal of generating responses suitable for new scenes with a small amount of training corpus [16].

1) Sequence-to-Sequence Model: The basic idea of an end-to-end dialogue system established using a sequence-to-sequence model is to generate a response based on a given query sentence.

Ritter et al. [38] propose a probabilistic model for generating responses. They collect large-scale corpus for training and build an end-to-end dialogue model based on a deep neural network to learn the correspondence between queries and responses. According to the trained dialogue model in the prediction stage, their method calculates the semantic vector of the input user and generates a reasonable response.

Shang et al. [41] build a corpus based on Sina Weibo data and uses the seq2seq model to realise the essential functions of a single-round dialogue system. The model uses two decoders: (1) a global decoder calculates the global semantic variables, and (2) a local decoder uses is the attention mechanism to weight the local semantic variables . By training such a sub-model, the two are combined and optimised, to maximise the actual recovery probability in training set for a given query. Shao et al. [42] take part of the generated sequence as an influencing factor in the attention mechanism.

That is, the sequence output by the model is divided into continuous segments of fixed length but not overlapping, and reconstructed into new training data. When predicting the model, they use the cluster searches in the segments and use the normalisation method to reorder the segments to generate responses. In order to improve the diversity of generated responses, Wu et al. [61] make improvements at the decoding stage based on a seq2seq model and propose another seq2seq model based on a dynamically decoded dictionary. In the training phase, the model learns the construction of a dynamic dictionary and the generation of responses. In the prediction phase, it dynamically provides a small-capacity vocabulary for the input questions and uses the vocabulary for decoding.

Mei et al. [31] integrate the RNN language model and the attention mechanism to achieve a single round of response generation system. The word sequence is input into the model to calculate the vector value of the hidden layer. At the decoding stage, each time the hidden vector of the previous time is used with an attention mechanism to calculate a weighted context vector. This vector combines with the hidden vector of the RNN language model calculation output. By providing different inputs for each generation step as a reference, they improve the diversity of generated responses. Li et al. [25] add an intention network based on the basic seq2seq model, which used to remember the user's intention and historical information in the dialogue. Wang et al. [54] analysis the dialogue corpus. They propose the method of integrating deep channel and full channel with a seq2seq model according to the phenomenon that the probability of the information in the context appearing in the response does not exceed 50%. The wideband channel introduces an attention mechanism to predict the keywords related to the input, and the deep channel training multilayer perceptron selects the keywords already in the input. The two are integrated with the input utterance and sent to the decoder that uses the attention mechanism to generate a response that is strongly related to the input utterance. Serban et al. [40] propose a latent variable-based hierarchical encoder-decoder model, which introduces a random variable in the context-based RNN. The latent variable class could be sentiment or other topics, which can be used to classify the responses. In the decoder, sample the latent variable first, and then generate the response. Due to the uncertainty of the latent variable, it can guarantee the rich and diverse content of the response.

Li et al. [28] propose a hierarchical annotation scheme and an end-to-end neural response generation model, using intent and semantic slots as intermediate sentence representations. They also design a filter based on whether or not these intermediate representations are suitable for the design task and dialogue constraints. An appropriate response is chosen, which can guide the user to complete the task while maintaining user participation. Wang et al. [53] propose a novel knowledge-aware response generation model. The model transfers question representation and knowledge matching capabilities from the knowledge base of question answering tasks to promote discourse power understanding and factual knowledge selection. Besides, they propose a response-oriented attention mechanism to enhance the encoder's understanding of the input and improve the knowledge selection through multi-step decoding to generate a more appropriate and meaningful response.

Compared with the non-hierarchical framework model, the hierarchical framework model can directly model the relationship at the dialogue level, enhances the dependency between query and response, and generates more diverse and meaningful responses. In future work, it is worth solving the problem of generating inconsistent responses to remote conversation history by developing methods that can effectively track longer dialogue history.

2) Reinforcement Learning: Researchers mainly use Reinforcement Learning (RL) [69] to solve sequence decision-making problems. It assumes that a series of the corresponding action outputs for a series of environmental changes, and then there will be a reward after the execution of these actions. The characteristic of RL is to learn the best strategy according to the environmental status, actions, and rewards. It does not see the benefits brought by the current action, but see the value that this action can bring in the future in order to obtain the maximum accumulation. To maximise reward is the goal to learn strategies. RL finds the best strategy through continuous trial and error, feedback, and learning.

Some researchers have used RL for the response generation of dialogue systems. Ilievski et al. [15] successfully apply transfer learning and deep RL to task-based chat robots. Due to the similarity between the source and target fields, in the environment of limited data, transfer learning compensates for the missing data. In the future, it is worth studying continuously how to use transfer learning in a dialogue system. Li et al. [24] propose a RL based model for response generation. They use the seq2seq model to learn the semantic relationships in the dialogue, and use RL to optimise the long-term dialogue mechanism. Song et al. [47] propose a RL based method for the generation of consistent responses for the dialogue system to human users. In particular, they use the natural language inference technology to find a reward signal for response generation. Koo et al. [19] integrate RL with attention-based layered recursive codecs and propose a personalised response generation system. First, it integrates user-specific information into the decoder to capture the user's background and speaking style information. Then it uses RL to maximise the return of future conversations so that the system can produce a coherent and natural response. Yang et al. [64] propose to use hierarchical RL for response generation. It learns abstract states or actions and decides which high-priority strategy should be selected. In the future, it is worth studying how to improve the design of different levels of strategy.

V. Hybrid Methods Based on Knowledge and Deep Learning

For some complicated questions, it is often impossible to retrieve the matched entries in a pre-set knowledge base. Thus a generation method is often required to generate responses to unmatched or meaningless questions. So it is necessary to integrate knowledge based retrieval methods and learning

based generative methods in order to have their strengths. This section will discuss some hybrid methods based knowledge and deep learning.

Li [26] proposes a hybrid model based on retrieval and response generation. The retrieval model indexes the question and response pairs in the corpus, retrieve the most similar questions by searching questions by the question and takes the response of similar questions as the response of the current question. The reordering method based on seq2seq sorts the response candidates in ascending order. The system takes top 1 as the response of the retrieval model before selecting. If the confidence score corresponding to the response from the retrieval model is higher than a pre-set threshold, the response is taken as the final response and returned to the user; otherwise, the final response is directly generated through the generation model based on seq2seq and returned to the user.

He et al. [12] propose an end-to-end Q&A system in sequence-to-sequence learning, which can answer complex questions. The Q&A system can dynamically predict the semantic units (words, phrases, and entities) in the generated response from the vocabulary, copied from the given question and jointly retrieved from the corresponding knowledge base, and finally, in the encoder-decoder generate responses to user questions within the framework.

Yan et al. [65] propose a retrieval-based dialogue system that can respond to different dialogue modes through a data-driven deep neural network framework. They combine the influence of multiple data inputs to model the context in multiple continuous dialogue interactions and incorporate valuable information into the query list through a retrieval mechanism. They optimise the integration of various functions and factors into the deep learning framework and argue that open-domain dialogue systems can adopt a standard dialogue model in different scenarios.

Qiu et al. [36] integrate the seq2seq model, a retrieval model and generative method for response generation. It uses the seq2seq model based on the attention mechanism to reorder the response candidates from the retrieval model First, the retrieval model is used to retrieve k candidates of Q&A pairs from the knowledge base of Q&A, and then use the rerank model to calculate the matching degree of each candidate of question and answer pair. If the score of the searched response candidate is not less than the pre-set threshold, it is regarded as the final response; otherwise, using the generative model to generate a new response. The Q&A knowledge base is constructed by extracting question and answer pairs from the online live user service log. However, as the number of real users increases, it is not easy to train suitable generative modules and construct reasonable retrieval models.

Tanaka et al. [50] propose a memory-enhanced hierarchical recursive encoder/decoder (MHRED), which is based on multi-round conversational context and external knowledge and is used to generate more diverse and meaningful responses. This model integrates the generated module, the search-based module and the reordering module. First, the response of the dialogue system is generated by MHRED and retrieved from a predefined database focused on fact retrieval. Then use the reordering module to sort these candidates according to the manually set rules, and finally select the highest score as the response. Therefore, the system can return diverse and meaningful responses from various angles. In the fire, it is worth introducing end-to-end learning for multiple systems at the same time, significantly improving the response generated by the dialogue system based on the combination of multiple modules.

Madotto et al. [30] propose a neural system model that enhances memory function. This model contains the conversation history and commonly used conversation knowledge stored in external storage components to minimize the lack of common sense and no logical response generated by the dialogue system.

Song et al. [48] use the results of the retrieval model as the input to the encoder in the seq2seq model. The results generated by the model are placed in the original candidate set and reordered. After such a process, the response given by the model is optimised.

Serban et al. [39] integrate natural language generation models and retrieval models, including template-based models, bag-of-words models, sequence-to-sequence neural networks, and hidden variable neural network models. They apply RL to crowdsourcing data and interacting with users to choose the appropriate response from its aggregated model.

Wang et al. [53] propose a multi-step decoding method, which can capture the knowledge connection between a question and the generated response to the question. In the first step, the response generated by the decoder and the draft response matches the relevant facts in the knowledge base. It makes the final response generated by the decoder in the second step more reasonable than the response generated in the first step. Also, they propose a response-oriented attention mechanism to enhance the encoder's understanding of user input problems and improve knowledge selection through multi-step decoding to generate a more appropriate and meaningful response.

Zhang et al. [67] propose a knowledge-aware attentive wasserstein adversarial response generation model. The model can can model external knowledge effectively and dialogue context-specific significant variances specific to dialogue interaction in a unified adversarial encoder/decoder learning framework. The universal attention module used to calculate the attention matrix and fuse the conversation context and response to train a better autoencoder. Besides, they proposed a novel knowledge-aware condition Wasserstein autoencoder to adjust the response based on the words in the conversation context and the external knowledge related to the conversation history, which can model explicit conversation semantics and implicit in a unified network architecture. Common sense. In plans, it is necessary to integrate more abundant knowledge into the dialogue system framework to generate more basic responses and apply the proposed model to various goal-oriented dialogue systems.

Xu et al. [62] propose an active dialogue generation model based on the knowledge graph, which generates an effective response in a limited knowledge graph. The model has three components: an improved model agnostic meta-learning algorithm (MAML), knowledge selection in knowledge ternary

embedding, and a knowledge-aware active response generator. MAML learns general features even when there are few entity relationships contained in the knowledge graph, and quickly perceives new knowledge to complete the dialogue generation task. The embedding and selection of knowledge triples are expressed as sentence embedding to better capture semantic information. The model's active learning ability can be applied to a learning system with knowledge reasoning to generate more meaningful responses.

Zhao et al. [70] in order to overcome the difficulty of lack of dialogue corpus, the decoder used for dialogue generation are decomposed into independent components, reducing the dependence on training data. The way to achieve this is to learn the central part of the model from a large number of unfounded dialogues and unstructured documents, and use the limited training examples to fit the remaining small parameters well. By separating the parameters that depend on the knowledge-based dialogue from the entire generative model, they can overcome the difficulties caused by insufficient training data. In the future, considering the lack of data, integrating more external knowledge will help the dialogue system to generate better responses.

Li et al. [27] realised an intelligent question answering system based on NBA basketball knowledge graph. The system automatically extracts the entities contained in the questions raised by the user and maps the extracted entities and the questions of the user to the corresponding entity attributes in the basketball knowledge graph and combines the retrieval mechanism to return the results to the user. Regarding the extraction of entities and the connection between multiple entities, it occupies a crucial position in the interaction of the dialogue system. How to accurately find out the relationship between multiple entities and combine detailed knowledge in multi-level relationship query It is a direction for future research on dialogue systems. The application of the knowledge base in the dialogue generation module can help the dialogue system generate a more reasonable and natural response. How to design the scope involved in the knowledge base, and how to integrate the knowledge related to retrieval and user input into the combination of various dialogue generation models, needs further research.

VI. Conclusion

As a service model of next-generation human-computer interaction, the dialogue system has received extensive attention from the academic industry and industry. The development of deep learning technology and the arrival of the era of big data have also brought new opportunities and challenges to data-driven dialogue systems. In this paper, we survey various response generation methods, from the oldest ones to the mainstream ones at the moment. In particular, we discuss various response generation methods based on deep learning, focusing on their problems and possible solutions to the problems.

In this survey, we attempt to establish a more complete the blueprint for building the response generation module of dialogue systems. Generally speaking, the effect of the dialogue model can be improved by improving the encoder, improving the decoder, and introducing the attention mechanism. Based on the reinforcement learning framework, the seq2seq model can work as a feedback mechanism to guide and improve the response gratin through rewards. Deep learning models are more and more widely in response generation, but the quality of the generated responses cannot be guaranteed, and even grammatical errors may occur. In the case of a limited data set, it will create unlimited single responses, such as good, I don't know and the lack of semantic information. These problems reduce users' satisfaction with dialogue systems. Combining external knowledge with the corpus of dialogue training is a way to make up for the gap in background knowledge between the dialogue system and the user. When introducing external knowledge into the dialogue system, the retrieval method will get a better matching performance, and the generative method will provide rich information.

References

[1] K. Abe, K. Kurokawa, K. Taketa, S. Ohno, and H. Fujisaki. A new method for dialogue management in an intelligent system for information retrieval. In *Processings of the 16th International Conference on Spoken Language Processing*, pages 1–4, 2008.

[2] J. Aron. How innovative is Apple's new voice assistant, Siri? *New Scientist*, 212(2836):24–24, 2011.

[3] C. Asakiewicz, E.A. Stohr, S. Mahajan, and L. Pandey. Building a cognitive application using watson deepqa. *IT Professional*, 19(4):36–44, 2017.

[4] A. Bordes, Y. L. Boureau, and J. Weston. Learning end-to-end goal-oriented dialog. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–15, 2017.

[5] T. Broens, S. Pokraev, J. Sinderen, M.V.and Koolwaaij, and P.D. Costa. Context-aware, ontology-based service discovery. In *Proceedings of the 2004 European Symposium on Ambient Intelligence*, pages 72–83, 2004.

[6] C. Chen, Q.Q. Zhu, R. Yan, and J.F. Liu. A summary of research on open domain dialogue system based on deep learning. *Chinese Journal of Computers*, 42(7):1439–1461, 2019. (In chinese).

[7] H. Chen, X. Liu, D. Yin, and J. Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

[8] B. Dhingra, L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 484–495, 2017.

[9] P. Ehrenbrink, S. Osman, and S. Möller. Google now is for the extraverted, cortana for the introverted: Investigating the influence of personality on ipa preference. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pages 257–265, 2017.

[10] M. Eric and C. D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, 2017.

[11] D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. A form-based dialogue manager for spoken language applications. In *Proceedings of the 4th International Conference on Spoken Language Processing*, pages 701–704, 1996.

[12] S.Z. He, C. Liu, K. Liu, and J. Zhao. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, page 199208, 2017.

[13] T. Holstein, M. Wallmyr, J. Wietzke, and R. Land. *Current Challenges in Compositing Heterogeneous User Interfaces for Automotive Purposes*, pages 531–542. Computer Science, 2015.

[14] L. Hurtado, J. Planells, E. Segarra, and E. Sanchis. Spoken dialog systems based on online generated stochastic finite-state transducers. *Speech Communication*, 83:81–93, 2016.

[15] V. Ilievski, C. Musat, A. Hossmann, and M. Baeriswyl. Goal-oriented chatbot dialog management bootstrapping with transfer learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence Organization*, pages 4115–4120, 2018.

[16] E.S. Juliano, F. Andre, and Siegfried H. An open vocabulary semantic parser for end-user programming using natural language. In *Proceedings of the 12th IEEE International Conference on Semantic Computing*, pages 77–83, 2019.

[17] K. Kim, C. Lee, D. Lee, J. Choi, S. Jung, and G.G. Lee. Modeling confirmations for example-based dialog management. In *Proceedingd of 3rd IEEE Spoken Language Technology Workshop*, pages 324–329, 2010.

[18] S. Kim, I. Kang, and N. Kwak. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 6586–6593, 2019.

[19] S. Koo, G.G. Lee, and H. Yu. Mathematical model for processing multi-user requests on POMDP hybrid dialog management. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, pages 1–4, 2016.

[20] J.-P. Kruth, T.V. Ginderachter, P.-I. Tanaya, and P. Valckenaers. The use of finite state machines for task-based machine tool control. *Computers in Industry*, 46(3):247–258, 2001.

[21] C. Lee, Y.S. Cha, and T.Y. Kuc. Implementation of dialogue system for intelligent service robots. In *Processings of 2rd International Conference on Control, Automation and Systems*, pages 2038–2041, 2008.

[22] C. Lee, S. Jung, M. Jeong, and G.G. Lee. Chat and goal-oriented dialog together: a unified example-based architecture for multi-domain dialog management. In *Proceedings of the 1st IEEE Spoken Language Technology Workshop*, pages 194–197, 2006.

[23] C. Lee, S. Jung, S. Kim, and G.G Lee. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, 51(5):466–484, 2009.

[24] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, 2016.

[25] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. In *Proceedings of the 22nd Empirical Methods in Natural Language Processing*, page 21572169, 2017.

[26] X.-S Li. Design and implementation of question answering system based on retrieval and answer generation. Master's thesis, 2019.

[27] Y. Li, J. Cao, and Y.B. Wang. Implementation of intelligent question answering system based on basketball knowledge graph. In *Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference*, pages 2601–2604, 2019.

[28] Y. Li, K. Qian, W.Y. Shi, and Z. Yu. End-to-end trainable non-collaborative dialog system. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 8293–8302, 2020.

[29] Z.D. Lu and H. Li. A deep architecture for matching short texts. In *Proceedings of the 2013 Neural Information Processing Systems*, page 13671375, 2013.

[30] A. Madotto, C.S. Wu, and P. Fung. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1468–1478, 2018.

[31] H. Mei, M. Bansal, and M.R. Walter. Coherent dialogue with attention-based language models. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3252–3258, 2017.

[32] F. Mi, M. Huang, J. Zhang, and B. Faltings. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence Organization*, pages 3151–3157, 2019.

[33] H. Noh, S. Ryu, D. Lee, K. Lee, C. Lee, and G.G Lee. An example-based approach to ranking multiple dialog states for flexible dialog management. *IEEE Journal of Selected Topics in Signal Processing*, 6(8):943–958, 2012.

[34] H.J. Oh, C.H. Lee, M.G. Jang, and K.Y. Lee. An intelligent TV interface based on statistical dialogue management. *IEEE Transactions on Consumer Electronics*, 53(4):1602–1607, 2007.

[35] M.-J. Peng, Y.W. Qin, C.X. Tang, and X.M. Deng. An e-commerce customer service robot based on intention recognition model. *Journal of Electronic Commerce in Organizations*, 14(1):34–44, 2016.

[36] M. Qiu, F.-L Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and Chu W. Alime chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, page 498503, 2017.

[37] A. Raux and M. Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 629–637, 2009.

[38] A. Ritter, C. Cherry, and W.B. Dolan. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, page 583593, 2011.

[39] I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, and N. R. Ke. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*, 2017.

[40] I.V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Aaron Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 2157–2169, 2017.

[41] L.F. Shang, Z.D. Lu, and H. Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 15771586, 2015.

[42] Y. Shao, S. Gouws, D. Britz, A. Goldie, B. Strope, and R. Kurzweil. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 22102219, 2017.

[43] B. Shen and D. Inkpen. Speech intent recognition for robots. In *Proceedingds of the 3rd International Conference on Mathematics and Computers in Sciences and in Industry*, pages 185–189, 2017.

[44] Y.L. Shen, X.D. He, L. Gao, J.F. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 2014 International World Wide Web Conference*, page 373374, 2014.

[45] Z.X. Shi and M.L. Huang. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 7007–7013, 2019.

[46] O. Sihombing, N. Zendrato, Y. Laia, M. Nababan, D. Sitanggang, W. Purba, D. Batubara, S. Aisyah, E. Indra, and S. Siregar. Smart home design for electronic devices monitoring based wireless gateway network using cisco packet tracer. *Journal of Physics Conference Series*, 1007(1):12–21, 2018.

[47] H.Y. Song, W.-N. Zhang, J.-W. Hu, and T. Liu. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 1–8, 2020.

[48] Y. Song, R. Yan, X. Li, D. Zhao, and M. Zhang. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*, 2016.

[49] Y.P. Song, X.Y. Zhou, and H. Wu. shall i be your chat companion?: Towards an online human-computer conversation system. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, page 649658, 2016.

[50] R. Tanaka, A. Ozeki, S. Kato, and A. Lee. Context and knowledge aware conversational model and system combination for grounded response generation. *Computer Speech & Language*, 62:1–10, 2020.

[51] A. Verma and A. Arora. Reflexive hybrid approach to provide precise answer of user desired frequently asked question. In *Proceedings of the 7th International Conference on Cloud Computing, Data Science and Engineering - Confluence*, pages 159–162, 2017.

[52] S.X. Wan, Y.Y. Lan, J.F. Guo, L. Xu, J. Pang, and X.Q. Cheng. A deep architecture for semantic matching with multiple positional sentence representations. In *Proceedings of the 2016 National Conference on Artificial Intelligence*, page 28352841, 2016.

[53] J. Wang, J.H. Liu, W. Bi, X.J. Liu, K.J. He, R.F. Xu, and M. Yang. Improving knowledge-aware dialogue generation via knowledge base question answering. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 1–8, 2020.

[54] W. Wang, M. Huang, X. Xu, F. Shen, and L. Nie. Chat more: Deepening and widening the chatting topic via a deep model. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, page 255264, 2018.

[55] Y. Wang, F.-J. Ren, and C.-Q. Quan. Review of dialogue management methods in spoken dialogue system. *Computer Science*, 42(6):1–6, 2015. (In chinese).

[56] Y.-G Wei, X.-M. Zhu, S. Bo, and B. Sun. Comparative studies of aiml. In *Proceedings of the 3rd International Conference on Systems and Informatics*, pages 344–349, 2016.
[57] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 438–449, 2017.
[58] J. Wu, M. Li, and C.H. Lee. A probabilistic framework for representing dialog systems and entropy-based dialog management through dynamic stochastic state evolution. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 23(11):2026–2035, 2015.
[59] Y. Wu, G. Nong, W.-H Chan, and L.-B. Han. Checking big suffix and lcp arrays by probabilistic methods. *IEEE Transactions on Computers*, 65(10):1667–1674, 2017.
[60] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z.J. Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, page 496505, 2017.
[61] Y. Wu, W. Wu, D. Yang, C. Xu, and Z. Li. Neural response generation with dynamic vocabularies. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5594–5601, 2018.
[62] H. Xu, J. Bao, and J. Wang. Knowledge-graph based proactive dialogue generation with improved meta-learning. *arXiv preprint arXiv:2004.08798*, 2020.
[63] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li. Building task-oriented dialogue systems for online shopping. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 4618–4625, 2017.
[64] M. Yang, Q.G Jiang, Y. Shen, Q.Y. Wu, Z. Zhao, and W. Zhoue. Hierarchical human-like strategy for aspect-level sentiment classification with sentiment linguistic knowledge and reinforcement learning. *Neural Networks*, 117:240–248, 2019.
[65] YanR., Y.P. Song, and H. Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, page 5564, 2016.
[66] W.P. Yin, H. Schtze, B. Xiang, and B. Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4(1):259–272, 2016.
[67] Y.Y. Zhang, Q. Fang, S.S. Qian, and C.S. Xu. Knowledge-aware attentive wasserstein adversarial dialogue response generation. *ACM Transactions on Intelligent Systems and Technology*, 11(4):1–15, 2020.
[68] Z.S. Zhang, J.T. Li, P.F. Zhu, H. Zhao, and G.S. Liu. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, page 37403752, 2018.
[69] T.-C. Zhao, K. Xie, and M. Eskenazi. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 12081218, 2019.
[70] X.L. Zhao, W. Wu, C.Y. Tao, C. Xu, D.Y. Zhao, and R. Yan. Low-resource knowledge-grounded dialogue generation. In *Proceedings of the 2020 International Conference on Learning Representations*, pages 1–14, 2020.
[71] X.Y. Zhou, L. Li, D.X. Dong, Y. Liu, Y. Chen, W.X. Zhao, D. H. Yu, and H. Wu. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, page 11181127, 2018.

Yifan Fan is currently a master student at Guangxi Normal University, China.

Dr. Xudong Luo is currently a distinguished professor of Artificial Intelligence at Guangxi Normal University, China. He published one book and more than 160 papers including 2 in top journal *Artificial Intelligence*, one of which has been highly cited by, for example, MIT, Oxford, and CMU research groups. Prof. Luo has international recognised reputation: co-chair and (senior) members of PC of more than 100 international conferences or workshops, including major conferences IJCAI and AAMAS, and referees for many international journals such as top journal *Artificial Intelligence*. He is also invited to make a presentation of his work in more than 10 universities internationally, including Imperial College. His research focus is on the areas of agent-based computing, fuzzy sets and systems, decision theory, game theory, knowledge engineering, and natural language process. Prof. Luo has supervised or co-supervised more than 40 master students, Ph.D. students, and research fellows.

Pingping Lin is currently a master student at Guangxi Normal University, China.