Neurocomputing 465 (2021) 248–264

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Towards information-rich, logical dialogue systems with knowledge-enhanced neural models

Hao Wang [a], Bin Guo [a,*], Wei Wu [b], Sicong Liu [a], Zhiwen Yu [a]

[a] Northwestern Polytechnical University, Xi'an, China
[b] Microsoft Corporation, Beijing, China

ARTICLE INFO

Article history:
Received 25 December 2020
Revised 16 July 2021
Accepted 29 August 2021
Available online 7 September 2021
Communicated by Zidong Wang

Keywords:
Text generation
Dialogue systems
Knowledge graphs
Neural network models

ABSTRACT

Dialogue systems have made massive promising progress contributed by deep learning techniques and have been widely applied in our life. However, existing end-to-end neural models suffer from the problem of tending to generate uninformative and generic responses because they cannot ground dialogue context with background knowledge. In order to solve this problem, many researchers begin to consider combining external knowledge in dialogue systems, namely knowledge-enhanced dialogue systems. The challenges of knowledge-enhanced dialogue systems include how to select the appropriate knowledge from large-scale knowledge bases, how to read and understand extracted knowledge, and how to integrate knowledge into responses generation process. Combined with external knowledge, dialogue systems can deeply understand the dialogue context, and generate more informative and logical responses. This survey gives a comprehensive review of knowledge-enhanced dialogue systems, summarizes research progress to solve these challenges and proposes some open issues and research directions.

1. Introduction

Text generation, also known as natural language generation (NLG), aims to make machines express like humans, which has the capability to produce smooth, meaningful and informative textual contents. From the original template-based and statistical methods to the deep learning-based methods, text generation has attracted the attention of massive researchers and made many remarkable advances [1]. The dialogue system is a typical application of text generation. Researchers have been committed to building dialogue systems that can communicate with humans naturally. Achieving truly fluent and natural human–machine conversation is a long-standing goal of artificial intelligence. In recent years, with the development of science and technology, more and more dialogue systems are flooding around us, making our life more convenient and comfortable.

According to different application fields, dialog systems can be divided into task-oriented dialog systems and non-task-oriented dialog systems. The former mainly helps users to complete specific tasks, such as restaurant reservation and travel time arrangement. Microsoft Cortana,[1] Google Assistant,[2] Amazon Alexa[3] and Apple Siri[4] are all typical task-oriented dialogue systems, which can assist us to complete various tasks in the form of conversations to reduce our operational burden. Non-task-oriented conversation systems, also known as chatbots, can communicate with users in the open domain. Chatbots do not help users complete specific tasks, but provide users with realistic and interactive dialogue experience and establish emotional connections with users. Microsoft Xiaoice[5] is the most popular and typical chatbot around the world, that has made conversations with hundreds of millions of users and successfully built long-time emotional connections with them.

With the development of deep learning, the researches of chatbots are also in a booming stage, and creating chatbots that can communicate naturally with humans is no longer a fantasy. With the help of the recurrent neural networks (RNN) [2], attention mechanism, generative adversarial networks (GAN) [3], reinforcement learning (RL), Variational Autoencoder (VAE) [4] and Transformer [5], dialogue systems have been able to generate smooth, topic-consistent and even personalized text. However, existing dialogue systems lack interactions with the real world, and have little access to the external knowledge, making them easily generate

* Corresponding author.
E-mail address: guob@nwpu.edu.cn (B. Guo).
[1] http://www.msxiaona.cn/.
[2] https://assistant.google.com/.
[3] https://developer.amazon.com/en-US/alexa.
[4] https://www.apple.com/siri/.
[5] http://www.msxiaobing.com/.

https://doi.org/10.1016/j.neucom.2021.08.131
0925-2312/

short and meaningless responses [6]. We humans are constantly acquiring, understanding and storing knowledge, and will automatically combine our knowledge to understand the current situation in communicating, which is a huge challenge faced by dialogue systems. To generate more informative, diverse and logical responses, dialogue systems must have the ability to combine external knowledge, which is a promising research direction.

Fig. 1 is an example of a dialogue system with or without commonsense knowledge, where we can see that combining with commonsense knowledge, including structured knowledge graph (KG) composed of knowledge triples and unstructured knowledge base (KB) composed of natural language text, the dialogue agent can have a deeper understanding of the dialogue context, and generate more informative and natural responses.

There are many researchers in both academia and industry have begun to explore knowledge-enhanced dialogue systems by incorporating different types of knowledge. Due to the complexity of the real world, the scale of external knowledge is usually extremely large. How to extract the most relevant knowledge from massive knowledge facts based on the relatively simple dialogue input is a huge challenge because of the diversity of natural languages. Meanwhile, how to effectively understand the extracted knowledge and integrate it into neural network models to facilitate response generation is also a difficult problem. This paper aims to give an in-depth survey of the development of knowledge-enhanced dialogue systems. To sum up, we summarize our contributions as follows.

- We briefly introduce the development process of dialogue systems, review the most widely explored neural network models in dialogue systems, and formalize the definition of knowledge-enhanced dialogue systems.
- We summarize the current research progress in knowledge-enhanced dialogue systems, including structured KG-enhanced dialogue systems and unstructured KB-enhanced dialogue systems by systematically categorizing the state-of-the-art works according to research challenges.
- We further summarize some commonly used datasets, benchmarks and evaluation metrics, and propose some open issues and research directions for the reference of the community, including combining structured and unstructured knowledge, lifelong learning, and so on.

The reminder of this paper is organized as follows. In Section 2, we give a brief review of key techniques applied in dialogue systems and in Section 3, we give the formalized definition of general dialogue systems and knowledge-enhanced dialogue systems. We then summarize the major research works of structured KG-enhanced dialogue systems and unstructured KB-enhanced dialogue systems in Sections 4 and 5, respectively. Finally, we summarize some commonly used datasets, benchmarks and evaluation metrics in this field in Section 6, and propose some promising research directions and conclude this paper in Section 7.

2. Key technologies

Deep learning has made great progress in many research fields, such as computer vision, speech recognition and natural language processing. Most advances in dialogue systems are also benefited from deep neural networks, mainly including RNN, GAN, RL, VAE, and Transformer. In this section, we will briefly introduce these technologies.

2.1. RNN

RNN owns sequential structure which is quite suitable for text sequence modeling. RNN processes sequential data in a sequential manner, and transforms the semantic information in the text sequence into the hidden state vector. Various variants of RNN are proposed, such as long short-term memory (LSTM), which can selectively process the important input information to the validity of sequence modeling by uniquely designed gating mechanisms. Based on LSTM, Sutskever et al. [7] propose the Sequence-to-Sequence (Seq2seq) framework, also known as the encoder-decoder framework. The encoder encodes the input sequence into semantic representations. The decoder generates words according to the semantic representations and previously generated words. Since it has no limitation of the length of input/output sequences, Seq2seq has become the standard architecture of dialogue systems.

2.2. VAE

VAE is one kind of generative model, composed an encoder which encodes the input sequence into latent variables, and a decoder to regenerate the original input based on the latent

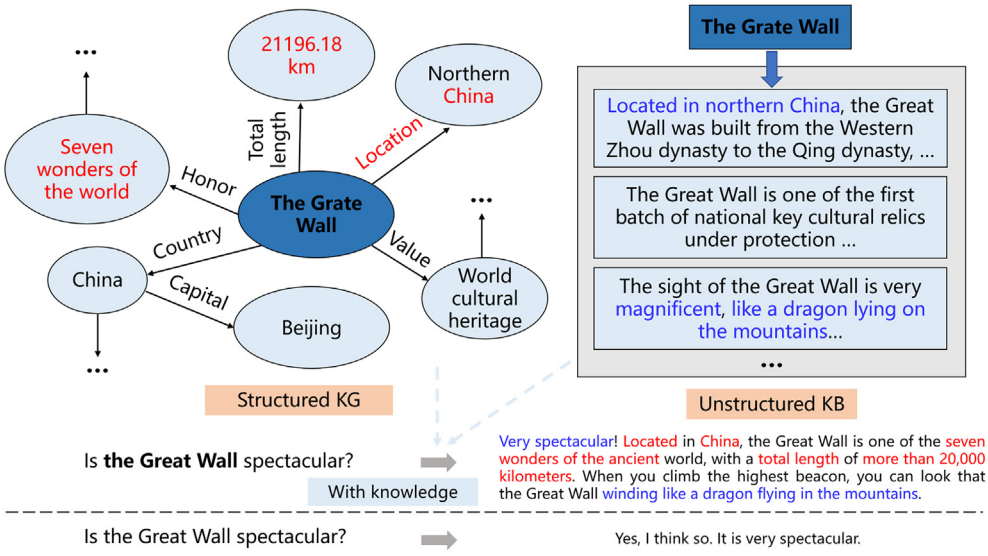


Fig. 1. Two dialogue examples with (the first line) or without (second) combining external knowledge.

variables. Given the input sequence X, the encoder encodes it into the latent space $P_\theta(z|X)$, where the z represents the latent variables and θ is the parameter of the encoder. Based on z, the decoder reconstructs the probability distribution of the input $P_\phi(X|z)$, where ϕ is the parameters of decoder. VAE can capture the latent hierarchical structure in natural language, such as topic and emotion, to generate more natural and informative dialogue responses.

2.3. GAN and RL

GAN, another powerful generative model, is also introduced to the researches of dialogue systems. There are two components in GAN, which are generator and discriminator, respectively. The generator generates false samples that are as close to the real data as possible, and the discriminator distinguishes the real data and the generated samples. With sufficient adversarial training, the generator is able to generate enough realistic samples, so as to imitate the human way of conversation and generate natural dialogue responses. RL is a Markov decision process, in which the agent takes actions to change its state and obtain reward (or punishment) according to the state. The goal of RL is to find the optimal policy and maximize the total rewards by trying various possible actions in different states. Combined with RL, GAN transforms the process of optimizing generator into the process of maximizing generator's reward. For example, *SeqGAN* [8] treat the text generation as a sequence decision process, where each generated word represents the state, the next word to be generated represents the action, and the discriminator's score of the generated text represents the rewards.

2.4. Transformer

Transformer [5] is a new sequence modeling model, composed of Attention modules and feed forward neural networks. The self-attention module encodes each word in a text sequence based on the context information, to obtain better semantic-rich vector representations of each word. Transformer has powerful semantic feature extraction capabilities and parallel computing capabilities, hence it has received extensive attention and make remarkable progress in many fields of natural language processing (NLP). The pre-trained language models, such as BERT [9] and GPT [10], mark the arrival of a new era of NLP research.

3. Knowledge-enhanced dialogue systems

With the help of the above neural network models and algorithms, the quality of generated dialogue responses has been greatly promoted. The development of dialogue systems requires the generated responses to be more informative, diverse and logical rather than simply smooth or superficially correct. Combined with external knowledge, dialogue systems can deeply understand the input, and make the generated responses more information-rich, more consistent with the logic of human expression, and with less common sense mistakes. Consideration of external knowledge in dialogue systems makes intelligent dialogue agent more anthropomorphic and bring better services for human beings in various fields. In this section, we will give the formalized definition of knowledge-enhanced dialogue systems.

We first give a formalized definition of general dialogue systems. Given the input dialogue query $X = (x_1, x_2, \ldots, x_m)$, the target of general dialogue systems is to generate the output dialogue response $Y = (y_1, y_2, \ldots, y_n)$, where m and n are the length of the query and response, respectively. The general dialogue systems can be defined as follows:

$$p(Y|X) = \prod_{t=1}^{n} p(y_t|X, y_{<t}) \tag{1}$$

At each decoding time step, the decoder will attend on the input query and the generated words in previous time steps to predict the current output word. After the sequential generation process, the dialogue response with the highest probability is generated. Based on the definition of general dialogue systems, we give the formalized definition of knowledge-enhanced dialogue systems.

Definition 1 (*Knowledge-enhanced dialogue systems*). Integrating external knowledge, such as commonsense or knowledge about specific events or concepts, to enhance the understand the conversation context, provide factual basis and reference for the dialogue process, and generate more informative, logical dialogue responses.

Given a set of knowledge facts $F = \{f_i\}_{i=1,2,\ldots,k}$, where each f_i may be a text sequence about specific concepts or a knowledge triple, and k is the number of facts. The knowledge-enhanced dialogue systems can be formulated as Eq. 2, where X and Y represent the input dialogue query and response, respectively.

$$p(Y|X, F) = \prod_{t=1}^{T} p(y_t|X, F, y_{<t}). \tag{2}$$

There are two forms of external knowledge, that is structured KG and unstructured KB. The KG is essentially a semantic network containing multiple types of entities and relations with the form of $< head, relation, tail >$, where *head* and *tail* are different entities. Entities refer to things in the real world and relations express connections between entities. The KB has no fixed form and usually store knowledge related to specific concepts in textual sequence form. Combined with external knowledge, dialogue agents can have a deeper understanding of dialogue context, make more reasonable reasoning, and generate more information-rich and logical dialogue responses.

There are many challenges in combining external knowledge into dialogue systems. When combining structured knowledge, the first challenge is how to obtain vector representations of knowledge triples as acceptable inputs to neural network models. Neural network models need input data with vector form, while the information stored in structured KB is symbolized. It is a difficult problem to map these symbols into low-dimensional dense vector spaces. And how to incorporate knowledge vectors as additional input into neural network models to guide the generation process is also a challenge. Because knowledge facts in the unstructured KB are stored with the form of natural language text, mapping knowledge facts to vector representations does not pose a research challenge. The first challenge in combining unstructured knowledge is how to select the most appropriate knowledge from massive knowledge facts due to the possible semantic duplication of different knowledge. The understanding of sentence-level natural language text is a long-term research challenge in NLP, so how to efficiently read and understand textual knowledge facts to integrate them into generation systems is another challenge in combining unstructured knowledge.

Fig. 2 is the visualized comparison of structured KG and unstructured KB-enhanced dialogue systems. The biggest difference between them lies in structures of knowledge facts, specifically the knowledge triples in KG and the knowledge texts in KB. In structured KG-enhanced dialogue systems, relevant knowledge triples can often be selected accurately by entity recognition methods, while transferring symbolic triples into vector representations by knowledge embedding is the crucial step in the encoder stage (as discussed in subsection 4.1). However, in unstructured

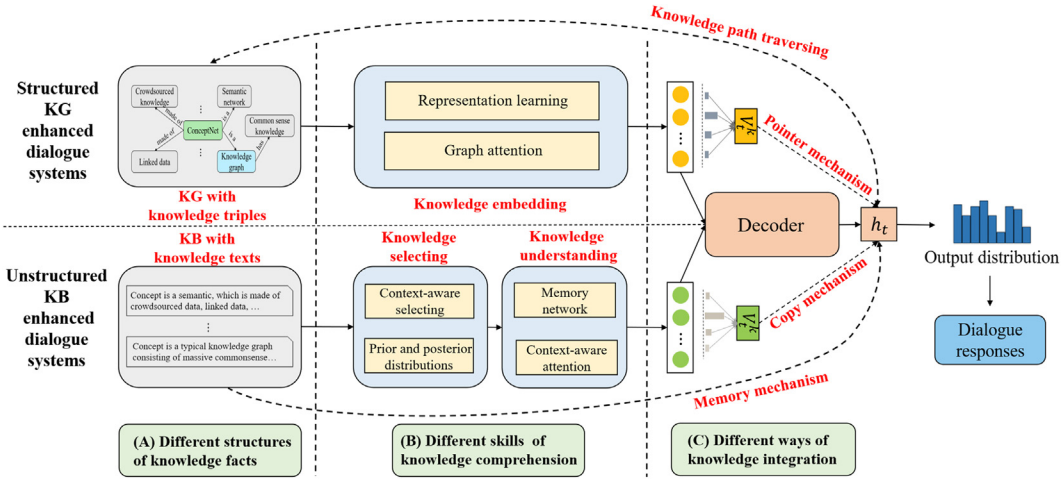


Fig. 2. The visualized comparison of structured KG and unstructured KB-enhanced dialogue systems.

KB-enhanced dialogue systems, knowledge selection is non-trivial because of the complex semantics of long natural language text. Therefore, the knowledge selection needs to be uniquely designed (as discussed in subsection 5.1) to provide appropriate knowledge for the knowledge understanding module (as discussed in subsection 5.2).

After encoding the dialogue context and relevant knowledge, the pointer and copy mechanism are adopted in structured KG and unstructured KB-enhanced dialogue systems to directly reference words from knowledge sources to generate more informative responses in the decoder stage (as discussed in subsections 4.2 and 5.2). Due to the graph structure, the structured KG often allows the graph path traversing to travel to more informative entities or attributes for response generation (as discussed in subsection 5.2). The memory mechanism is explored to store and interact with dialogue context to enhance the understanding of unstructured knowledge (as discussed in subsection 4.2). Researchers have made efforts to incorporate external knowledge into dialogue systems, which will be summarized in following sections.

4. Structured KG-enhanced dialogue systems

Structured KGs can store a wider range of knowledge types in many domains but relatively less information due to their simple representation of triples. However, its unique symbolic storage form is quite different from vectors required by neural networks. Therefore, how to map knowledge triples into low-dimensional vector representations (*i.e.*, knowledge embedding) and efficiently incorporate knowledge vectors into neural network dialogue models are key directions of the research. We give a brief summary of structured KG-enhanced dialogue systems in Table 1. We divide the research directions into structured knowledge embedding and knowledge graph incorporation, and summarize some novel research directions, which will be detailed introduced as follows.

The general framework of KG-enhanced dialogue systems is shown as Fig. 3. Given the dialogue context, the entity matching method is usually adopted to retrieve relevant knowledge triples. Then knowledge triples are embedded into vectors by key knowledge embedding techniques, including the sequence embedding (as discussed in subsection 4.1.1), the representation learning (as discussed in subsection 4.1.2) and the graph attention (as discussed in subsection 4.1.3). Subsequently, it is necessary for the dialogue decoder to incorporate knowledge into the decoding procedure to generate informative and logical dialogue responses. The knowledge incorporation methods mainly include the knowledge path traversing (as discussed in subsection 4.2.1), the pointer mechanism (as discussed in subsection 4.2.2), and the context-aware knowledge incorporation (as discussed in subsection 4.2.3).

Table 1
A summary of structured KG-enhanced dialogue systems.

Structured knowledge embedding	Embedding triples as word sequences	Young et al. [11], Yang et al. [12], Liu et al. [13]
	Representation learning	Zhou et al. [14], Moon et al. [15]
	Graph attention mechanism	Zhou et al. [14]
Knowledge graph incorporation	Knowledge path traversing	Moon et al. [15], Zhang et al. [16], June [17]
	Pointer-network mechanism	Madotto et al. [18], Wu et al. [19]
	Context-aware incorporation	Wu et al. [20], Wu et al. [21], Wu et al. [22], Yang et al. [23]
Novel directions	Reinforcement learning	Xu et al. [24], Xu et al. [25]
	Knowledge transferring	Wang et al. [26]
	Knowledge Disentangling	Raghu et al. [27], Madotto et al. [28]

4.1. Structured knowledge embedding

The symbolized form of knowledge triples in structured knowledge graphs makes it difficult to incorporate knowledge into dialogue systems. Semantic accurate vector representations of knowledge triples are the premise of rational use of relevant knowledge in dialogue systems. Researchers have proposed many methods to map knowledge triples into semantic vector representations, which will be summarized as follows.

4.1.1. Embedding knowledge triples as word sequences

The simplest and effective way to obtain vector representations of structured KGs is to directly treat entities and relations in knowledge triples as common word sequences, and encode them to obtain vector representations as the same as the dialogue context sequence encoding. After that, the knowledge vectors can be integrated into the decoder to generate more informative responses.

For instance, Young et al. [11] firstly integrate a large KG into dialogue systems. They extract triples from KG according to entities in the input query to form text sequences, which are encoded by LSTM in knowledge encoder to obtain representations. Then

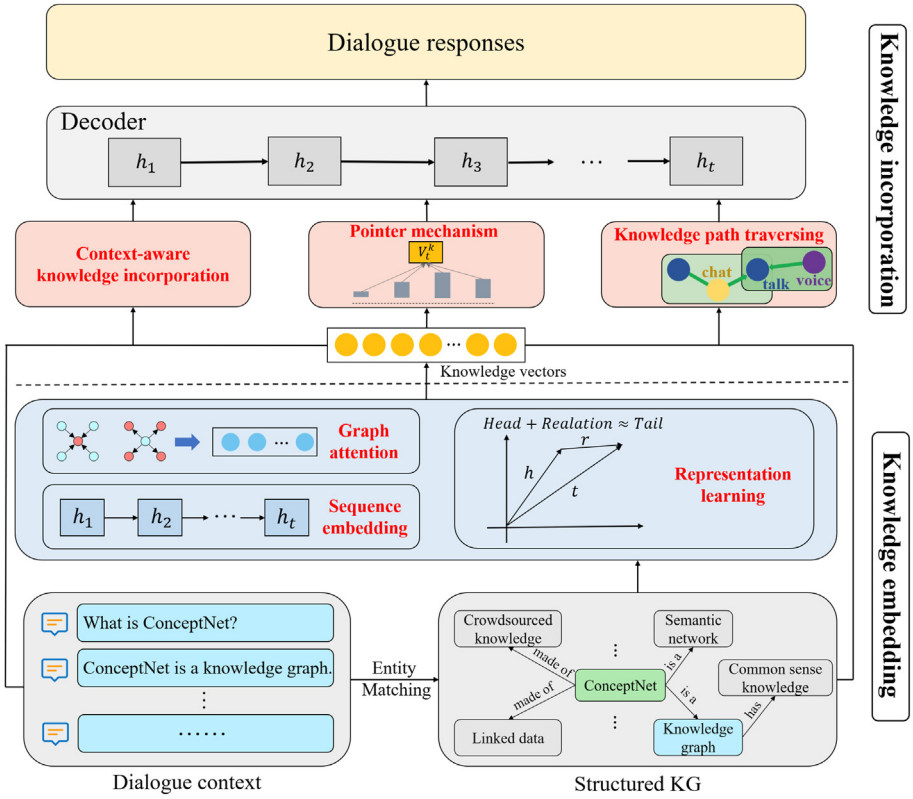


Fig. 3. The general framework of KG-enhanced dialogue systems.

knowledge vectors are added with the input vector to calculate the degree of correlation with the alternative responses to enhance the accuracy of the response selection. Similarly, Yang et al. [12] utilize LSTM to encode each knowledge triples into vector representations, and then summarize all relevant facts into a fixed-size vector by averaging their hidden vectors. The knowledge vectors are fed into both the generation model and the retrieval module to obtain high-quality candidate dialogue responses. The hybrid ranking module selects the best response among retrieved and generated response candidates. Considering that conversation usually drifts from one entity to another, Liu et al. [13] propose the entity diffusion to augment the dialogue model with the ability of convergent and divergent thinking over the knowledge base. The similarity between extracted entities and other entities is calculated to retrieve relevant entities. Word vectors of entities and relations are averaged to obtain vector representations of each triple, which is concatenated with the input vector to guide the response generation.

Embedding knowledge triples as word sequences can simply and directly obtain the knowledge guidance according to the keywords in the input text, for better understanding of the dialogue context. However, due to the symbolic characteristics and sparse expression of KGs, this approach loses a large amount of structural and connection information of the adjacent knowledge entities hidden in the graph, and fails to utilize KGs comprehensively.

4.1.2. Representation learning-based knowledge embedding

Representation learning aims to learn a series of low-dimensional dense vectors to represent semantic information, while knowledge representation learning is designed for entities and relations in the knowledge graph, to map them into vectors for performing subsequent calculations and inferences. To obtain the vector representations of symbolic knowledge in the knowledge graph, researchers propose many knowledge representation learning algorithms, in which the TranE algorithm [29] is the most classic algorithm. TransE is a distributed vector representation based on entities and relations. The relation in each triple, $< head, relation, tail >$, is regarded as the translation from head to tail entity, to update vector representations of knowledge triples.

For example, Zhou et al. [14] adopt the TranE algorithm to get the initial vectors representations of knowledge triples. They retrieve knowledge subgraph from the whole KG using the words in the input query as queries, and then transform each triple in the knowledge subgraph into vector representations using pre-trained TranE embeddings. Similarly, Moon et al. [15] construct KG embeddings to encode each entity in the KG using TranE algorithm. Rely on the structural information in the knowledge graph, TranE algorithm can obtain knowledge vector representations with rich semantic information. Then the knowledge vectors can be further enhance the understanding of the dialogue context and the response generation.

There are complex semantic and reasoning relationships between entities in KGs. The distributed vector representations of entities and relations obtained through representation learning can naturally integrate these semantic associations, thus providing more information for informative response generation. Meanwhile, representation learning can alleviate the problem of data sparseness in KGs, and integrate heterogeneous information in entities and relations. However, the large-scale knowledge graph makes the computational complexity of graph algorithms very high, and the long-tail distribution of data makes it difficult to capture the semantic information of a large number of entities and relations.

4.1.3. Graph attention-based knowledge embedding

To facilitate the usage of graph structure information in structured KG, the graph attention mechanism is introduced into the

KG-enhanced dialogue systems. Graph attention mechanism can obtain more efficient knowledge vector representations by considering the characteristics of adjacent entities and relations in knowledge graph, so as to enhance the performance of dialogue systems.

Zhou et al. [14] firstly propose the static and dynamic graph attention mechanism to facilitate dialogue context understanding and response generation using the large-scale commonsense knowledge graph, as shown in Fig. 4. In the encoder stage, the Knowledge Interpreter adopts the static graph attention to measure the association of a relation to a head entity and a tail entity for generating the static vector representation for a graph. The representation of the retrieved graph will be concatenated with the input vector to enhance the understanding of the dialogue context. And in the decoder stage, the Knowledge Aware Generator adopts dynamic graph attention to attentively read all knowledge graphs and then read all the triples in each graph for generating information-rich responses.

By considering not only entities but also relations in the KGs, the graph attention mechanism encodes more structured semantic information and the association information among nodes, which can obtain more semantically rich vector representations of the knowledge, for reasonable responses. Meanwhile, considering only local information of adjacent nodes and relations cannot effectively fuse the structure information of the whole graph, leading to a loss of information.

4.2. Knowledge graph incorporation

After obtaining the vector representations of knowledge triples, the key to improve the performance of KG-enhanced dialogue systems is to facilitate the understanding of the dialogue context based on the knowledge facts and reasonably incorporate them into the generated responses. Researchers have made various attempts to incorporate structured KG into response generation, which will be summarized as follows.

4.2.1. Knowledge path traversing for knowledge incorporation

According to some research on discourse development, human conversations are not "still", which means that we will chat around a series of related knowledge concepts and shift our focus from one concept to others as the conversation evolves. This phenomenon indicates that there exits walkable knowledge paths within a KG to form graph traversal steps, which lead to more informative entities or attributes for response generation.

In response to the above observations, Moon et al. [15] propose the DialKG Walker model to learn knowledge paths among entities mentioned over dialog contexts and reason over large scale KG. The graph decoder attends on several KG paths to predict the most appropriate knowledge facts according to the dialogue context and the facts ever been mentioned. Then the parallel zero-shot learning model transforms entities into embedding spaces to enhance the response generation. Similarly, Zhang et al. [16] introduce the ConceptFlow model to explicitly model conversation flows using commonsense KGs. As shown in Fig. 5(a), the concepts of a conversation evolve from "chat" and "future", to relevant concepts "talk" and "dream", along the commonsense relations. ConceptFlow constructs the concept graph based on the input dialogue context, and models the conversation flow as traverses guided by graph attentions to moving towards more informative directions to generate more high-quality and semantic responses, as shown in Fig. 5(b).

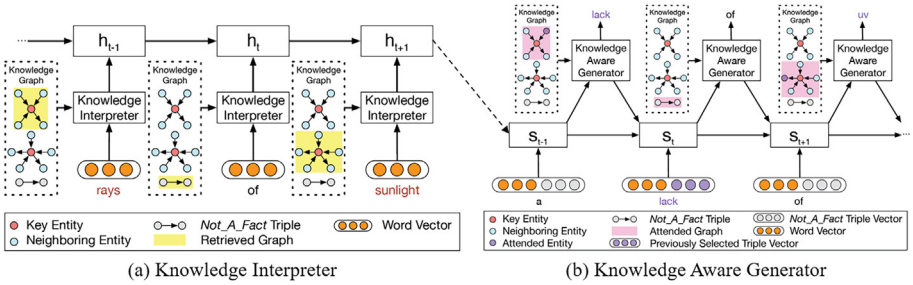
To better leverage KG structure information in KG path traversal for knowledge selection, June [17] propose the AttnIO model, composed of the incoming attention flow and outgoing attention flow to model the KG path traversal. By propagating the attention weights at each node to its neighbor nodes, the two attention flow explore KG to find out the most relevant entities about the dialogue context.

By mimicking the path traversing process of knowledge and topics in real human conversations, the dialogue system can model the conversation development along more meaningful directions in the KGs, to efficiently integrate knowledge information according to the dialogue context, for generating more informative and on-topic responses. Meanwhile, the path traversing process starting from a certain knowledge node can reduce the search space for relevant knowledge to improve the efficiency of knowledge utilization.

4.2.2. Pointer mechanism for knowledge incorporation

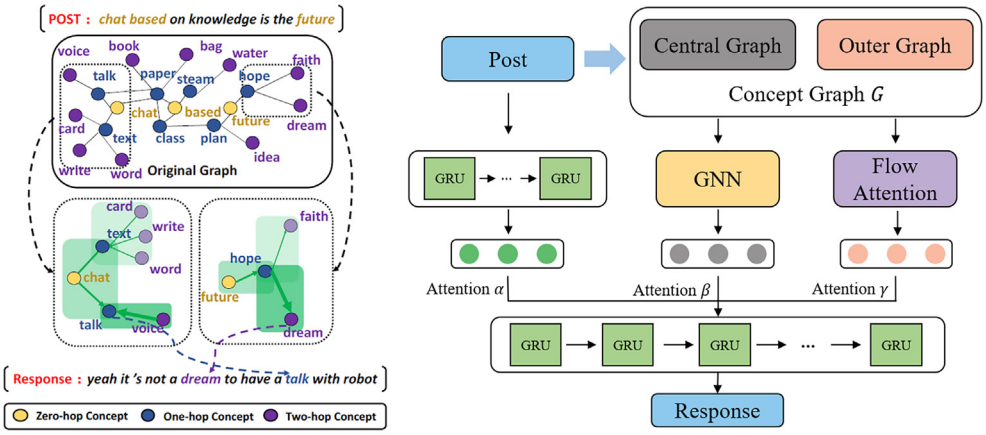
Because the scale of structured KG is usually extremely large, it is very hard to encode and decode the knowledge information in the KG due to the huge information noise, making the response generation unstable. To more explicitly incorporate external knowledge, the pointer networks [30] are introduced into the KG-enhanced dialogue systems, which can obtain the probability distribution under the input text, to realize the directly copy of essential words from input source to output responses.

Madotto et al. [18] firstly propose to combine the multi-hop attention over memories with the pointer network to effectively incorporate information in KG. The encoder create stored memories through semantic interaction of the dialogue context and relevant knowledge triples. Then the decoder generates words by pointing to the input words in the memory. At each decoding time step, two distributions are generated: one over all vocabulary words and one over the memory contents. The two distributions are combined to generate final responses. To improve the effectiveness of knowledge copy, Wu et al. [19] introduce the global-to-local memory pointer (GLMP) networks, composed the global memory and local memory. The global memory pointer softly filters unnecessary words for copying according the dialogue context, and the local memory firstly generates sketch responses and copy words from external knowledge to fulfil the sketch slot values.



(a) Knowledge Interpreter (b) Knowledge Aware Generator

Fig. 4. The model structure of commonsense conversational model (CCM) [14].



(a) The example of knowledge shift (b) The model structure of ConceptFlow

Fig. 5. The example of knowledge shift and model structure of ConceptFlow [16].

The pointer mechanism can make explicit use of external knowledge and dialogue context by copying knowledgeable words to enhance the response generation, rather than just using the fixed dimensional hidden states to encode conversation information. Meanwhile, the process of copying words reduces the complexity of generating long word sequences and avoids the OOV (Out Of Vocabulary) situation. However, the pointer mechanism also loses the variety of generated responses. How to balance the generation and copying is a key issue in the research.

4.2.3. Context-aware knowledge incorporation

Dialogue context is the key information to retrieve and understand external knowledge. To enhance the decoder's understanding of the retrieved knowledge information, it is necessary to fuse the dialogue context and the relevant knowledge to facilitate the integration of the knowledge in the KG-enhanced dialogue systems.

For instance, Wu et al. [20] design the Felicitous Fact mechanism to select knowledge facts highly relevant to the dialogue context under the guidance of human-generated answers as the posterior context knowledge, as shown in Fig. 6. The Felicitous Fact Recognizer reads the contextual information, and then outputs a probability distribution over knowledge facts to detect the facts that highly coincide with the dialogue context. The Context-Knowledge Fusion fuses the dialogue context and retrieved knowledge together based on the knowledge distribution with the Bag-of-Words Loss to ensure the accuracy of the input. The Triple Knowledge Decoder generates three types of words: copied words, knowledge words and vocabulary words, with the decision of Flexible Mode Fusion. Wu et al. [21] propose to fine-grained rank the retrieved knowledge based on the context information to facilitate knowledge selection. The Knowledge Bridge Fusion module adopts cross attention mechanism to fuse the semantic information both in the dialogue context and the knowledge to enhanced the informative response generation.

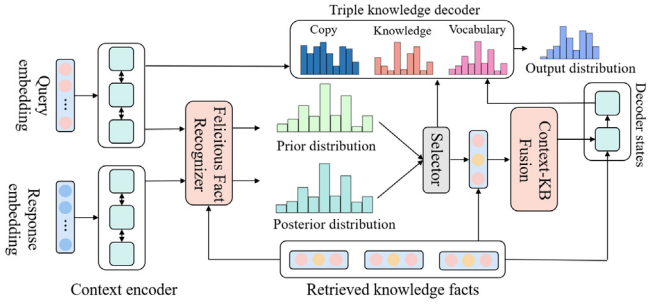


Fig. 6. The model structure of context-aware knowledge incorporation in [20].

Topic information is an important information in the conversation context, which can guide the flow direction of the conversation. To explicitly utilize knowledge and make the generation process more controllable and interpretable, Wu et al. [22] introduce the TopicKA model to generate dialogue responses conditioned both on the dialogue context and a interpretable topic fact with an explicit semantic meaning. The topic facts are commonsense facts which are semantically and logically connected with the dialogue context and the target response, and continuously involved during the conversation to guide the direction of conversation. Yang et al. [23] investigate the GraphDialog model to exploit the graph knowledge both in dialogue context and KGs into task-oriented dialogue systems. The multiple hidden states obtained by the masked attention mechanism in novel designed recurrent unit encode the dialogue context into graph structural representations. The multi-hop reasoning on the KG aims to retrieve relevant entities to improve the model consistency.

The human conversations are very complex, and different dialogue context makes the utilized knowledge various. Therefore, utilizing dialogue context to understand and integrate external knowledge in the response generation is of great help to improve the performance of KG-enhanced dialogue systems.

4.3. Novel directions for KG-enhanced dialogue systems

In addition to the above research works, researchers consider many other ways to enhance the performance of the KG-enhanced dialogue systems, which will be briefly summarized below.

4.3.1. RL-based KG-enhanced dialogue systems

RL is the basic framework for universal artificial intelligence. Through uniquely designed reward mechanisms, RL agents can learn to find the optimal solution in the large sample space. RL has been used to encourage coherent, informative, and long-lasting utterance sequences, so as to make the dialogue process more engaging. It is a promising research direction to combine RL to model the conversation process and integrate knowledge into KG-enhanced dialogue systems more reasonably.

Xu et al. [24] introduce the RL mechanism into KG-enhanced dialogue systems to seamlessly incorporate rich medical knowl-

edge graph and symptom-disease relations into the topic transition in Dialogue Management. The Knowledge-routed Deep Q-network (KR-DQN) (a typical network of RL) integrates a knowledge-routed graph branch and a relational refinement branch to make full use of knowledge of doctor experience. The knowledge-routed graph branch enhances the policy decision through a well-designed medical knowledge graph routing prior information based on conditional probabilities to make more informative and reasonable dialogue decisions. Considering that there are always topic changes in human conversation, Xu et al. [25] divide the dialogue conversation task into two sub-tasks, that are explicit goal sequence planning and goal completion by topic elaboration, respectively. They propose a three-layer RL-based KnowHRL model to complete two tasks. The upper-layer policy learns to traverse a KG to complete a high-level conversation goal, the middle-layer and lower-layer generate dialogue response about the specific topic under the goal-driven generation mechanism. The goal-sequence planning assigns the dialogue agent to conduct active conversations towards recommended topics, which brings more personified conversation experience.

With the uniquely designed reward mechanisms, RL can constrain the response generation at a higher semantic level (e.g., information richness, fluency), resulting in higher quality responses. However, due to the abstractness of natural language semantics, it is quite difficult to reasonable design these language-related rewards to control the generation process.

4.3.2. Cross-task knowledge transferring

Different NLP tasks have some similar characteristics, and have their own areas of expertise. For example, the knowledge base question answering (KBQA) system aims to accurately match the target knowledge triples from structured KG according to the input question and generate the corresponding answer of the given question. Meanwhile, the KG-enhanced dialogue systems need to find out relevant knowledge triples from KG to enhance the response generation. Therefore, it has important research value to transfer the task knowledge of different NLP tasks, so as to support the knowledge-enhanced dialogue systems to better select and understand knowledge, and to generate more high-quality responses.

For example, Wang et al. [26] propose the TransDG model, which transfers the question representation and knowledge matching abilities of KBQA models to effectively fuse external knowledge in an external KG for generating informative dialogue responses. The pre-trained KBQA model is composed of an encoding layer to obtain the vector representations of input query and candidate answers, and a knowledge selection layer to accurately select relevant knowledge entities. Then the pre-trained encoding layer and knowledge selection layer are combined into the original KG-enhanced dialogue model to transfer the utterance understanding and knowledge selection ability to generate more appropriate and informative responses.

Knowledge in other tasks can provide a priori guidance for KG-enhanced dialogue systems, making the knowledge selection and incorporation more effective. However, due to the heterogeneity of different tasks, knowledge transfer may bring a certain degree of noise, resulting in performance loss.

4.3.3. Disentangling knowledge from conversation

The size of external KGs is usually extremely large, with a huge number of entities and relationships, making the end-to-end dialogue systems hard to scale when incorporating knowledge facts in KGs. Therefore, researchers propose to disentangle the knowledge representation and incorporation from the traditional end-to-end generation process, to facilitate knowledge incorporation in the dialogue systems.

For example, Raghu et al. [27] propose the Bag-of-Sequences (BOSS) memory to disentangle the language model and its knowledge incorporation in response generation to deal with changes in external KGs. The higher level flat memory encodes triples and dialogue contexts and the lower level encoding of each individual utterance and tuple is constructed by GRU to disentangle the language and knowledge during response generation. Instead of directly using KG as input, Madotto et al. [28] choose to embed any size of KB into the dialogue model parameters. The proposed Knowledge Embedded (KE) approach consists of a user goal query that generates equivalents KE dialogues from the KG to store the information of KG into the model parameters to generate correct and informative responses. Disentangling knowledge from conversation can isolate parameters that are relevant to knowledge-ground dialogues from the entire model, to reduce the dependence of model of knowledge-grounded dialogues and improve the generalization of KG-enhanced dialogue systems.

5. Unstructured KB-enhanced dialogue systems

Unstructured KBs are composed of natural language text related to various concepts, which express rich semantic information. Because of its textual form, the unstructured KB can be easily combined with text generation systems whose inputs are text sequences. However, due to the information redundancy caused by a large amount of knowledge text, how to extract the most appropriate knowledge from massive knowledge facts is a big challenge. Meanwhile, the understanding of sentence-level natural language text is a long-term challenge in NLP, so how to efficiently read and understand textual knowledge facts to integrate them into generation systems is another challenge in combining KBs. We give a brief summary of unstructured KB-enhanced dialogue systems in Table 2. We divide the research directions of unstructured KB-enhanced dialogue systems into appropriate knowledge selection and knowledge understanding and integration, and summarize some novel research directions, which will be detailed introduced as follows.

The general framework of KB-enhanced dialogue systems is shown as Fig. 7. Given the dialogue context, the first step is selecting appropriate knowledge facts from a large number of knowledge texts by prior and posterior distributions-facilitated knowledge selection (as discussed in subsection 5.1.1) or context-aware knowledge selection (as discussed in subsection 5.1.2). Then different technologies are adopted to understand and incorporate the knowledge facts into the decoding state to generate more meaningful dialogue responses, including the memory mechanism (as discussed in subsection 5.2.1), the copy mechanism (as discussed in subsection 5.2.2), and Transformer-based models (as discussed in subsection 5.2.3).

5.1. Appropriate knowledge selection

Knowledge selection is a critical step in KB-enhanced dialogue systems which can extremely influence the model learning to make full use of the knowledge. Researchers have explored various methods to select appropriate knowledge, which will be summarized as follows.

5.1.1. Prior and posterior distributions-facilitated knowledge selection

In the process of knowledge selection of dialogue system, the most direct idea is to select specific knowledge based on the semantic similarity between the input query and the knowledge. This kind of semantic similarity is defined as the *prior distribution over knowledge*. However, there are one-to-many phenomena between the query and the knowledge, which means that different

Table 2
A summary of unstructured KB-enhanced dialogue systems.

Appropriate knowledge selection	Prior and posterior distributions	Lian et al. [31], Chen et al. [32], Kim et al. [33]
	Context-aware knowledge selection	Zhang et al. [34], Ren et al. [35], Zheng et al. [36], Meng et al. [37]
Knowledge understanding and integration	Memory mechanism	Ghazvininejad et al. [6], Dinan et al. [38]
	Copy mechanism	Yavuz et al. [39], Meng et al. [40], Lin et al. [41]
	Transformer	Zhao et al. [42], Li et al. [43], Zhao et al. [44]
Novel directions	Knowledge transferring	Qin et al. [45], Tian et al. [46]
	Pre-trained language model	Wang et al. [47], Zhao et al. [48], Tuan et al. [49]
	Few-shot learning	Zhao et al. [50]

knowledge can be used to generate different responses according to the same input query. In contrast, the *posterior distribution over knowledge*, which is inferred from both the input query and the corresponding response, can provide better guidance for knowledge selection. During the inference process of the dialogue system, there are no actual responses, so how to use the prior distribution to enhance posterior knowledge selection is the focus of research.

Lian et al. [31] notice that the discrepancy between the prior and posterior distributions over knowledge brings difficulties to accurate knowledge selection. They obtain the prior and posterior distributions over knowledge through attending on input query, and the combination of input query and target response, respectively, as shown in Fig. 8. By approximating the posterior with the prior distribution through minimizing the Kullback–Leibler divergence loss (KLDivLoss) between them, the prior distribution can be utilized to select appropriate knowledge, so as to generate informative responses even the actual response is unknown. Similarly, Chen et al. [32] consider reducing the gap between prior and posterior knowledge selection in KB-enhanced dialogue systems. The prior knowledge selection is enhanced by the necessary posterior information and the Knowledge Distillation Based Training Strategy (KDBTS) trains the decoder with prior knowledge, removing the exposure bias of knowledge selection. Since we will choose different knowledge in multi-turn conversation, Kim et al. [33] leverage a sequential latent variable model to model the knowledge selection in dialogue systems. The sequential knowledge transformer model keeps track of prior and posterior distribution over knowledge and is dynamically updated according to the previous selected knowledge and responses to select the most appropriate knowledge for response generation.

Under the guidance of certain responses, the posterior distribution over knowledge can help the dialogue systems to reasonably select the relevant knowledge, thus improving the performance of KB-enhanced dialogue systems. However, in the model inference, it is very difficult to accurately approximate the posterior distribution without corresponding responses, and the wrong approximating can lead to inappropriate knowledge selection.

5.1.2. Context-aware knowledge selection

In the multi-turn dialogue process, the conversation context may influence the knowledge selection of the following dialogue turns. Therefore, it is necessary to modeling context information in the knowledge selection process. For example, Zhang et al. [34] introduce the Context-aware Knowledge Pre-selection (CaKe) to leverage the dialogue context as a query to search a set of positions of the knowledge and select the most relevant knowledge under the modified bi-attention flow mechanism. Ren et al. [35]

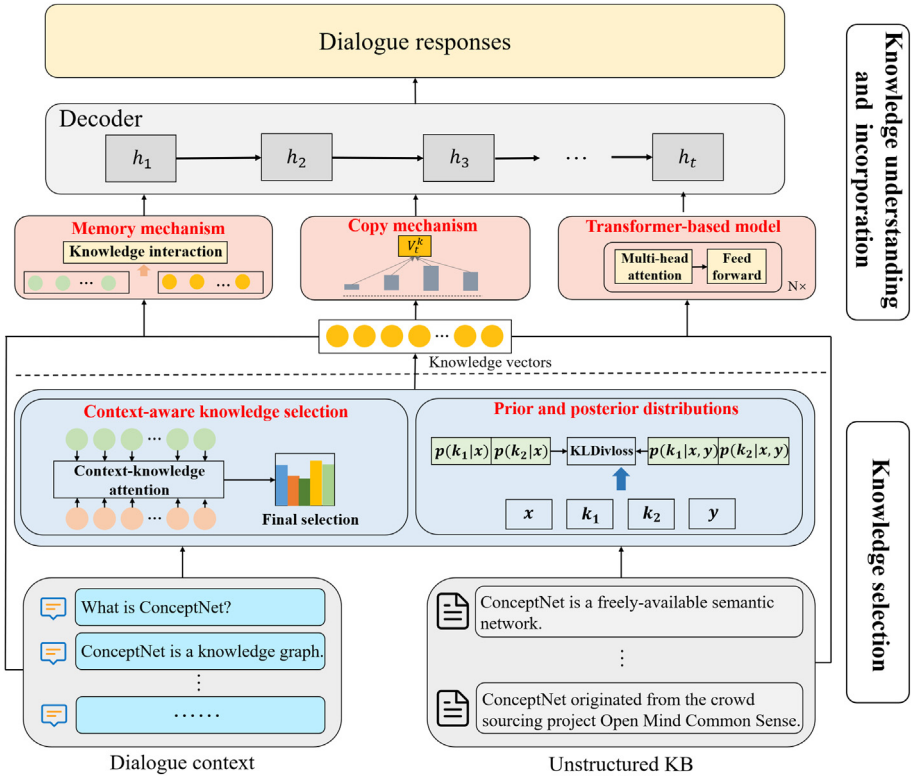


Fig. 7. The general framework of KB-enhanced dialogue systems.

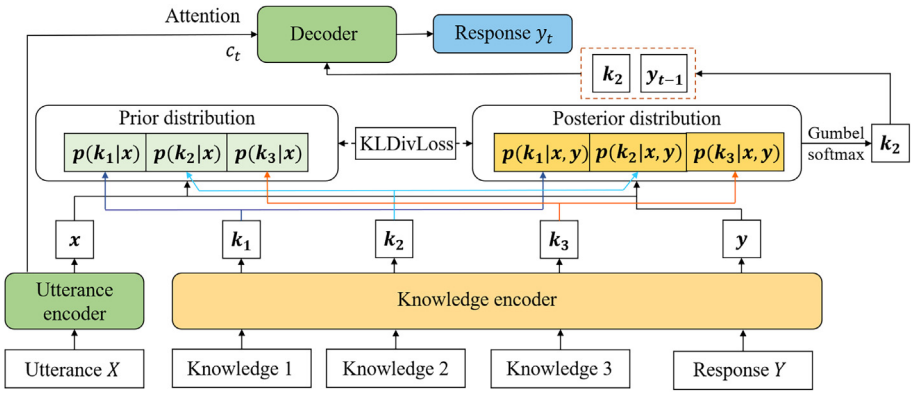


Fig. 8. The prior and posterior distributions over knowledge [31].

propose a Global-to-Local Knowledge Selection mechanism using the global perspective to select appropriate background knowledge. A topic transition vector is learned from the dialogue context and external knowledge by a distantly supervised learning schema to select the most likely text fragments. The vector is then used to guide the Local Knowledge Selection module at decoding stage to generate fluency and appropriate responses.

To model the difference of knowledge selection in multi-turn conversation, Zheng et al. [36] propose DiffKS to compute the difference between the candidate knowledge facts to be selected at this turn and previous selected knowledge and then the difference informative is fused with or disentangled from the dialogue context to facilitate the accurate knowledge selection. Meng et al. [37] explicitly model knowledge tracking and knowledge shifting as dual tasks and propose an unsupervised learning scheme to facilitate interactions between knowledge tracking and knowledge shifting to select more appropriate knowledge in response generation process.

The continuity of human multi-turn conversations leads to the continuity and correlation of knowledge selection. Selecting relevant knowledge according to the dialogue context can effectively improve the accuracy of knowledge selection, thus improving the informativeness of generated responses.

5.2. Knowledge understanding and integration

How to read and understand the selected appropriate textual knowledge to enhance the dialogue context representation and integrate it into the response generation process is the key problem in KB-enhanced dialogue systems.

5.2.1. Memory mechanism for knowledge understanding

Memory networks [51] are proposed to improve the poor memory ability of RNN, using external memory component to realize the storage of long-term memory. End-to-end memory networks [52] realize the end-to-end training process by reducing the transmission of supervised signals. Memory networks can store the selected external knowledge and interact with the dialogue context to enhance the understanding of both the knowledge and dialogue context and obtain the semantically richer vector representations of them. Memory networks are widely used in KB-enhanced dialogue systems to retrieve, read and condition on external knowledge text.

For instance, Ghazvininejad et al. [6] propose to store retrieved knowledge facts according to keywords in the input query into memory networks. Then the facts encoder attends on the knowledge facts based on the input query to get the semantic representations of knowledge, which are fed into the decoder along with the input to generate informative responses. To carefully read and understand the retrieved knowledge, Dinan et al. [38] combine memory networks and Transformer to encode the selected knowledge and the dialogue context to get the higher level semantic representation. Then the dot-product attention between the knowledge and context is performed to retrieved most relevant knowledge for generating the next response.

The memory module in memory networks can store and understand the knowledge with long textual form more effectively through explicit storage and recursive information interaction, thus facilitating the understanding of dialogue context and response generation. However, when the scale of external knowledge is very large, the computational complexity of memory networks increases sharply, leading to the low efficiency of knowledge understanding.

5.2.2. Copy mechanism for knowledge integration

General dialogue systems are mostly based on the encoder-decoder framework. The encoder encodes the input query into the hidden vector representation, which is fed into the decoder to generate the dialogue response. This process cannot explicitly utilize external knowledge facts, so the copy mechanism is introduced into the encoder-decoder framework, which enables the model not only to generate words in the vocabulary, but also to copy the appropriate knowledge words from the input, so as to realize the explicit integration of the external knowledge in the response generation.

For example, Yavuz et al. [39] propose the DEEPCOPY model that extends the Seq2seq model with a hierarchical pointer network to enable the decoder to copy words from either external knowledge facts or dialogue context, as shown in Fig. 9. Two copy probabilities, responsible for copying words from the dialogue context and each knowledge fact, respectively, are combined by the inter-source meta attention to obtain the final copy probability. Meanwhile, a soft switch mechanism smoothly combines the generation and copying distributions in the response generation process to generate both fluent and informative dialogue responses. Traditional copy mechanism can only copy one word at a time from external knowledge facts, which makes the dialogue systems relatively ineffective in leveraging background information. To better locating the right relevant knowledge, Meng et al. [40] incorporate the reference decoder that can directly copy entire semantic unit (e.g., a span expressing complete semantic information) from external knowledge facts to generate more informative responses. Lin et al. [41] propose the recurrent knowledge interaction to dynamically select different knowledge along with decoding steps to facilitate multiple knowledge into generated responses. Then the knowledge-aware pointer network is designed to copy words

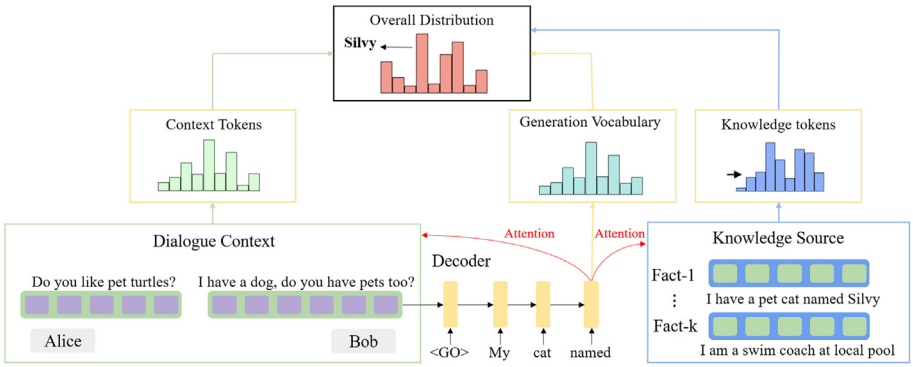


Fig. 9. The model structure of DEEPCOPY [39].

from external knowledge facts to generate more fluent, coherent and informative responses.

5.2.3. Transformer-based knowledge understanding and integration

Transformer is an emerging sequential model, which has caused great influence in NLP. It exceeds RNN in semantic information abstraction, long-term feature extraction, and task comprehensive feature representation. Based on the self-attention mechanism, the knowledge facts can be comprehensively understand through the interaction process with the dialogue context to obtain the semantically rich representations, and further attended to be integrated into the response generation.

For example, Zhao et al. [42] make use of multi-head attention mechanism in Transformer to encode the dialogue context, response candidate and the relevant knowledge document. Through the hierarchical interaction in the context and document, the importance of different parts of the document and context is determined to select the most appropriate response. Li et al. [43] propose the Incremental Transformer with Deliberation Decoder for multi-turn document grounded conversations, as shown in Fig. 10. The Incremental Transformer encodes multi-turn dialogue context with knowledge under an incremental encoding scheme. The previous utterances u^i and the relevant knowledge documents s^i are fed into the Incremental Transformer to obtain the vector representations incrementally with the attention mechanism. The Deliberation Decoder contains two processes where the first-pass focuses on contextual coherence and the second-pass refines the results of the first-pass by attending on the knowledge to increase the knowledge relevance and correctness.

To fuse different kinds of knowledge into dialogue systems, Zhao et al. [44] propose a universal Transformer-based architecture, composed of the encoder responsible for encoding dialogue context enhanced by multifarious knowledge, and the decoder responsible for generating informative responses with the knowledge-aware mechanism.

The powerful language understanding and modeling capabilities of Transformer enable it to efficiently read and understand knowledge in long textual form, and effectively integrate the knowledge with the dialogue context, to generate informative responses. However, when the length of knowledge text is longer, the computational complexity of Transformer will increase sharply, resulting in a decrease of computational efficiency.

5.3. Novel directions for KB-enhanced dialogue systems

In addition to the above research works, researchers consider many other novel ways to enhance the performance of the KB-enhanced dialogue systems, which will be briefly summarized below.

5.3.1. Cross-task knowledge transferring

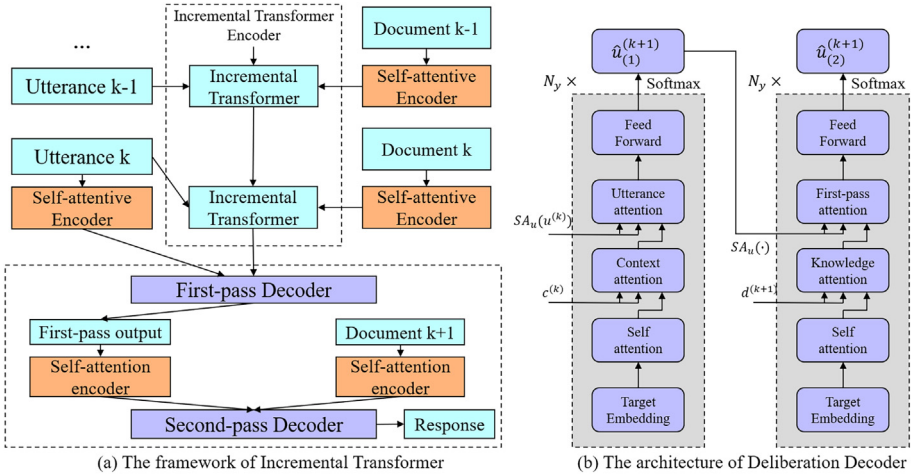
As mentioned in the section of KG-enhanced dialogue systems, other NLP tasks can also enhance the performance of KB-enhanced dialogue systems. For example, reading comprehension models are specialized in comprehensively understanding the long text documents to find out appropriate answers of the given question. Meanwhile, KB-enhanced dialogue systems need to read on the long textual knowledge to enhance the understanding of dialogue context and the response generation. Therefore, transferring the ability of other NLP task models is a promising way to improve the performance of KB-enhanced dialogue systems.

To enhance the ability of knowledge integration, Qin et al. [45] propose to jointly learn response generation together with on-demand machine reading. Combined with the state-of-the-art reading comprehension model, the external knowledge is projected as the document in the reading comprehension to present notably richer and more complex information for more informativeness and diversity response generation. Considering the difference between response generation and reading comprehension, Tian et al. [46] construct a response-anticipated memory to contain import document information to generate more informative responses. The teacher model constructs a weight matrix to contain information about the importance of words in knowledge based on the target response, and the student model learns to mimic the weight matrix without the access of response. After fully trained, the student model has the ability to select the most relevant knowledge to enhance the quality of generated responses.

5.3.2. Pre-trained language model-enhanced dialogue systems

The idea of pre-training has been widely explored in many NLP fields in recent years. By pre-training the language model on massive text corpus to initialize most of the network parameters which learn universal knowledge about syntactic and semantic information of neural language, and fine-tuning the model using a small amount of specific downstream task data, excellent performance can be achieved. Many pre-training models based on Transformer, such as BERT [9], GPT [10], RoBERTa [53], ELECTRA [54], and UniLM [55], have proved that pre-trained language models can greatly promote the performance of various NLP tasks. Various works have also been done to explore the pre-training language models in KB-enhanced dialogue systems.

Wang et al. [47] leverage the pre-trained BERT model as the encoder to encode the dialogue context and external knowledge text to obtain the more semantic-rich vector representations. Then the transformer decoder fuses the dialogue context and knowledge representations in sequential or joint way to generate informative responses. Zhao et al. [48] focus on equipping the pre-trained response generation model with a knowledge selection module which retains key information relevant to the dialogue context.



(a) The framework of Incremental Transformer (b) The architecture of Deliberation Decoder

Fig. 10. The model structure of Incremental Transformer with Deliberation Decoder [43].

The learning of knowledge selection and fine-tuning of responses generation are in an unsupervised manner, without requiring labeled dialogues with selected knowledge to release the model from human annotations dialogue data. Similarly, to solve the data problem, Tuan et al. [49] propose the Decoupling model to train KB-enhanced dialogue systems without paired external knowledge data. The Decoupling first de-couples knowledge-related information from the dialogue input and then use the pre-trained language model as the knowledge base to extract information from knowledge facts. Finally, the model re-couples the knowledge back into the input to enhance the response generation.

The giant pre-trained language model learns the universal linguistic knowledge from a large number of natural language texts and memorizes enough patterns in language, which can provide a priori guidance for the understanding of dialogue context and relevant knowledge, thus improving the informativeness of generated responses and the generalization ability of the dialogue systems.

5.3.3. Few-shot learning

The training of deep neural network-based knowledge-enhanced dialogue systems needs massive paired KB-ground dialogue datasets. Although knowledge documents and text are very abundant on internet, it is difficult to obtain large-scale dialogue datasets that are naturally grounded on the knowledge documents for training a neural dialogue model. Some research institutions have released some KB-ground dialogue datasets, but it is still scarce for the extensive research. To solve this problem, few-shot learning is introduced into the research area, which can learn a high-quality KB-enhanced dialogue system with as few knowledge-grounded dialogues as possible.

Zhao et al. [50] make the attempt of low-resource KB-enhanced dialogue generation. They represent a disentangled response decoder to separate parameters relying on knowledge-grounded dialogues from the whole model to solve the problem of lacking knowledge-grounded training data, as shown in Fig. 11. The decoder is composed of three components, including Language Model to generate common words, Context Processor to generate context words, and Knowledge Processor to generate words from knowledge document by a hierarchical attention mechanism. The Decoding Manager dynamically determines which components are activated at each timestep for informative response generation. Only the knowledge processor and the decoding manager are depending on the knowledge-grounded dialogues. Since the parameters of them are small in scale, the whole model can achieve expected performance with only a small number of knowledge-grounded dialogues.

The few-shot learning can effectively reduce the demand of knowledge-ground dialogue data samples required by the training of large-scale neural network models, thus improving the generalization ability of KB-enhanced dialogue systems.

6. Datasets, benchmarks and evaluation metrics

The training of knowledge-enhanced dialogue systems needs the support of supervised learning benchmark tasks which exhibit knowledgeable open dialogue with clear knowledge grounding. To free the researchers from repetitively and difficultly data collection to train the dialogue model, many high-quality knowledge-grounded dialogue datasets have been released. Meanwhile, another key issue faced by researchers is how to effectively evaluate the performance of proposed knowledge-enhanced dialogue models. Only with the help of reasonable evaluation metrics can researchers precisely and fairly evaluate the performance of designed models. In this section, we will summarize the commonly used datasets, benchmarks and evaluation metrics in knowledge-enhanced dialogue systems.

6.1. Datasets and benchmarks

To make researchers focus on the algorithm design of the knowledge-enhanced dialogue systems, many institutions have released knowledge-grounded dialogue generation benchmark datasets, as shown in Table 3, where the Dialogs and Utters represent the total number of dialogues and utterances in the dataset, respectively.

HelloE.[6] HelloE [56] is a dialogue dataset containing conversations about movies, where in each dialogue response is explicitly generated by copying and/or modifying sentences from unstructured background knowledge such as plots, comments and reviews about the movie. The whole dataset contains around 90K utterances of 9K conversations about 921 movies, whose explicitly linked to external knowledge can facilitate the development of knowledge-enhanced dialogue systems.

CMU_DoG.[7] CMU_DoG [57] is a document grounded conversation dataset where each conversation is followed by specified documents about popular movies extracted from Wikipedia articles, to

[6] https://github.com/nikitacs16/Holl-E.
[7] https://github.com/festvox/datasets-CMU_DoG.

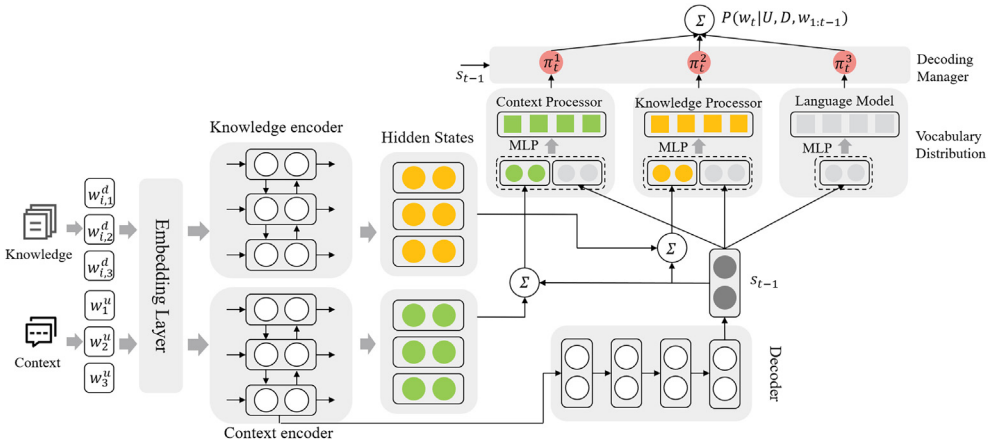


Fig. 11. The model structure of low-resource KB-enhanced dialogue generation [50].

Table 3
Comparison between different knowledge-grounded dialogue datasets.

Dataset	Domain	Knowledge type	Dialogs	Utters
Hello-E [56]	Movie	KB	9K	90K
CMU_DoG [57]	Movie	KB	4K	86K
WoW [38]	Open	KB	22K	202K
Opendialkg [15]	Open	KG	15K	91K
DyKGChat [58]	TV series	KG	1.2K	17K
Duconv [59]	Movie	KB&KG	30K	270K
KdConv [60]	Film, Music, Travel	KB&KG	4.5K	86K
MovieChats [61]	Movie	KB&KG	270K	3M

address the concerns of grounding in conversation responses, context and coherence in dialogue systems. The whole dataset contains around 4K conversations with an average of 21.43 turns per conversation.

Wizard of Wikipedia.[8] Wizard of Wikipedia (WoW) [38] is an open-domain dialogue benchmark for training dialogue agents that can converse knowledgably about open-domain topics. The whole dataset contains around 22K conversations which are directly grounded with knowledge retrieved from Wikipedia. The test set is split into two subsets, Test Seen and Test Unseen. Test Seen contains conversations whose topics are overlapped with the training set, while Test Unseen contains conversations whose topics are never seen before in train or validation.

OpenDialKG.[9] OpenDialKG [15] is a human-to-human multi-turn dialogs dataset where the mention of entities in each conversation is annotated in a large-scale common fact knowledge graph, containing 91K utterances across 15K dialog sessions. This dataset provides researchers a new way to study how conversational topics could jump across many different, grounded on KG paths that thread all of them.

DyKgChat.[10] DyKgChat [58] is a dialogue benchmark about how to apply dynamic knowledge graphs in neural conversation to facilitate dialogue systems to learn and zero-shot adapt zero-shot adapt with dynamic knowledge graph. This dataset is composed of TV series conversations of a Chinese palace drama "Hou Gong Zhen Huan Zhuan", containing around 1.2K dialogues and 17K dialogue turns.

DuConv.[11] DuConv [59] is a knowledge-driven conversation dataset where each conversation is grounded with factoid knowledge graph or unstructured non-factoid knowledge base under the given conversation topics. It contains around 270K sentences within 30K multi-turn conversations, and enables a very challenging task as the model needs to both understand dialogue and plan over the given knowledge graph.

KdConv.[12] KdConv [60] is a Chinese multi-domain knowledge-enhanced dialogue dataset, which grounds the topics in multi-turn conversations to external knowledge (structured knowledge triples or unstructured knowledge text). It contains 4.5K conversations from three domains (film, music, and travel), and 86K utterances, which carry in-depth discussions on related topics and natural transition between multiple topics.

MovieChats.[13] MovieChats [61] is a Chinese conversational corpus with fine-grained annotations in the movie domain, where each conversation is labeled with entity information and dialogue acts classified into 15 fine-grained aspects, based on which linked into different types of knowledge. It contains around 270k dialogues with over 3M utterances.

6.2. Evaluation metrics

There are mainly two kinds of evaluation metrics to evaluate the performance of knowledge-enhanced dialogue systems at the present stage, namely automated evaluation metrics and human evaluation metrics. Automatic machine evaluation metrics can automatically evaluate the performance of dialogue systems by comparing the difference between the responses generated by the dialogue systems and the ground-truth responses. The more similar the generated responses are to the target responses, the higher the quality of the dialogue systems. Human evaluation can evaluate the overall quality of the generated responses or at the finer-grained level, such as informativeness, relevance, knowl-

[8] https://parl.ai/projects/wizard_of_wikipedia/.
[9] https://github.com/facebookresearch/opendialkg.
[10] https://github.com/Pascalson/DyKGChat.
[11] https://ai.baidu.com/broad/subordinate?dataset=duconv.
[12] https://github.com/thu-coai/KdConv.
[13] https://github.com/chin-gyou/MovieChats.

edge utilization, etc. Due to these characteristics cannot be captured by algorithms, human evaluation is almost the only the most effective way to accurately evaluate the performance of knowledge-enhanced dialogue systems.

Perplexity. Perplexity [62] is a metric used to evaluate the quality of language models, and the uniform standards of all text generation models, including dialogue systems. The Perplexity represents the average number of uncertain tokens when the generative models predicting words and the smaller the num is, the better the model performance is. The fluency and diversity of generated dialogue responses can be reflected to some extent by Perplexity.

BLEU. BLEU [63] is the harmonic mean of *n*-gram precisions of generated responses and ground-truth responses. The *n*-gram precisions represent how many *n*-gram units in the generated responses are matched with those in the reference responses. The higher the BLEU score, the more similar the generated response is to the target response, and the higher the quality of dialogue systems.

ROUGE-L. Rouge-L [64] aims to calculate the longest common subsequences (*LCS*) between the generated responses and the ground-truth responses, with the same words in the same order. The *F*-measure is calculated based on the maximum precision and recall of reference responses to obtain the final ROUGE-L score, where the accuracy and recall are calculated by dividing the length of *LCS* by the length of the generated and reference response, respectively.

Distinct. Distinct [65] is designed to measure diversity of generated responses. The Distinct is calculated by dividing the number of unique *n*-gram units in the generated responses by the number of all *n*-gram units in the responses, and the score higher the score, the higher the diversity of generated responses.

7. Conclusion and future directions

This survey makes a comprehensive literature review of the research trends of knowledge-enhanced dialogue systems. Knowledge is the universal expression of all things and concepts in the real world and an important way to understand the real world. With the access to the external world knowledge, dialogue systems can comprehensively understand the dialogue context, think about how to reasonably respond, organize the logical expression, fuse the knowledge into the response, and finally generate information-rich, logical and diversified dialogue responses. The research of knowledge-enhanced dialogue systems not only has a broad application prospect but also facilitates dialogue agents to comprehensively understand human language and to be more anthropomorphic. Due to the syntactic and semantic complexity of natural language, the reasonable knowledge selection and integration is still great challenges for knowledge-enhanced dialogue systems. We review a variety of representative research efforts in terms of knowledge selection and integration in this survey. As an emerging research direction, there are still many open issues to be resolved, which will be briefly discussed as follows.

7.1. Combining structured and unstructured knowledge

At present, researches mainly focus on incorporating one form of external knowledge to enhance the performance of dialogue systems. Structured KG can narrow down knowledge candidates using the prior information such as entities and graph paths. Unstructured KB can provide abundant information to enhance text generation but we need strong capability of natural language understanding to select useful information. If the two forms of knowledge are combined, more appropriately and informatively dialogue responses may be generated. For example, Niu et al. [66] explore the possibility of combining them. Taking the commonsense KG as the backbone, the knowledge texts are aligned with the KG by linking entities from sentences to entities within the KG. The RL and the reading comprehension technology are combined to conduct fine-grained knowledge understanding and selection based on the dialogue context. Both forms of knowledge have their own advantages and disadvantages. Due to their structural differences, it is a challenging research direction to combine the structured and unstructured knowledge into generation systems, which will certainly bring promising progress to the knowledge-enhanced dialogue systems.

7.2. Knowledge extraction from crowdsourced data

The structured KG and unstructured KB are all needed to be organized manually to provide ground to the knowledge-enhanced dialogue systems. With the rapid development and popularization of socil networks, massive crowdsourced data appear on the Web, such as the Q&A community Quora[14] and Zhihu,[15] which reflects the implicit knowledge of human intelligence [67]. The existing knowledge-enhanced dialogue systems can only extract relevant knowledge from the given KG or KB, without the ability of performing real-time knowledge selection and fusion from crowdsourced data. Crowd intelligence data covers nearly every domains of our daily life and dynamically updates itself in real time. How to mine the appropriate informative from crowdsourced intelligence data containing massive noise to enhance the understanding of dialogue context and generation of responses is a promising research direction.

7.3. Pre-training technologies for dialogue systems

The emergence of pre-training has brought NLP to a new era, as discussed above. Various pre-trained language generation models (e.g., UniLM [55], GPT-3 [68]) have demonstrated their effectiveness on generative tasks with simple fine-tuning on a small amount of task-specific data. However, direct application of large-scale pre-training models to downstream tasks does not yield as many improvements as might be expected. For example, as discussed in the paper [69], directly applying fine-tuned GPT-2 on story generation task still suffers from generating repetitive, inconsistent and illogical stories, without the guidance of sufficient knowledge. Therefore, integrating knowledge into pre-training models is the promising research direction to release the superior performance of pre-trained models in various NLP tasks, which has been tentatively explored [70–72]. It is thus important to incorporate dialogue-related knowledge into pre-trained models to facilitate the understanding of the conversation context and the generation of responses of knowledge-enhanced dialogue systems.

7.4. Long-term memory ability

Multi-turn dialogue models require the ability to store long-term dialogue history and external knowledge. The dialogue context is essential for understanding the whole dialogue process and prompting the conversation conducted along the engaging directions. Meanwhile, the memory of external knowledge is important for understanding and integrating them into the dialogue models. However, existing neural network techniques, such as RNN and Transformer, have poor long-term memory storage capacity, which causes serious information loss in multi-turn con-

[14] https://www.quora.com/.
[15] https://www.zhihu.com/.

versations and reduces the coherence of conversations. Nevertheless, memory networks [51] try to utilize external memory component to store long-term memories and have been extensively researched in knowledge-enhanced dialogue systems [6,38]. How to memorize the long-term dialogue context and knowledge into the model structure itself, so as to achieve more efficient understanding of the conversation, is the key to achieve more information-rich and logical dialogue systems.

7.5. Meta learning for universal dialogue systems

The training of knowledge-enhanced dialogue systems requires extensively labelled dialogue data grounded on specific knowledge, which is unavailable in many dialogue domains. Although there have been various knowledge-grounded dialogue datasets released by some research institutions (e.g., Wizard of Wikipedia [38], KdConv [60]), open domain dialogue systems still face the data scarcity problem. Meanwhile, the existing KGs and KBs are often incomplete and covers only a few specific domains, making it extremely hard to train open domain dialogue systems with excellent performance. Meta-learning [73] is a potential research direction to address the above challenges, which utilizes previous knowledge and experience to guide the learning process of new tasks, so that the models have the ability to learn to learn. There have been some researches to apply meta learning to dialogue systems [74,75], which proves the ability of meta learning on adapting to new conversation domains with minimal training dialogue samples. It is very promising to combine meta learning algorithms with knowledge-enhanced dialogue systems, to adapt to the model smoothly on new conversation domains without enough domain knowledge, thus achieving truly intelligent dialogue systems.

7.6. Lifelong learning

We humans continuously learn new knowledge, update our knowledge base to adapt to the fast-changing pace of society. However, existing dialogue systems mostly utilize fixed knowledge bases whose knowledge do not keep updating in real time. To make text generation models more anthropomorphic, they should have the ability of continuous lifelong learning. A meaningful exploration of this is discussed by Mazumder et al. [76]. They propose the lifelong interactive learning and inference model which will actively ask users questions when encountering unknown concepts, and update its knowledge base after corresponding answers are reached. How to continuously obtain information from numerous external inputs and achieve lifelong learning is a important research direction in knowledge dialogue systems.

CRediT authorship contribution statement

Hao Wang: Writing - review & editing. **Bin Guo:** Writing - review & editing. **Wei Wu:** Writing - review & editing. **Sicong Liu:** Writing - review & editing. **Zhiwen Yu:** Writing - review & editing.

Declaration of Competing Interest

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

Acknowledgements

This work was partially supported by The National Science Fund for Distinguished Young Scholars (62025205), and The National Natural Science Foundation of China (Nos. 61772428, 61725205).

References

[1] B. Guo, H. Wang, Y. Ding, W. Wu, S. Hao, Y. Sun, Z. Yu, Conditional text generation for harmonious human-machine interaction, ACM Transactions on Intelligent Systems and Technology (TIST) 12 (2) (2021) 1–50.
[2] J.L. Elman, Finding structure in time, Cognitive Science 14 (2) (1990) 179–211.
[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680. .
[4] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013). .
[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008. .
[6] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-T. Yih, M. Galley, A knowledge-grounded neural conversation model, in: AAAI, 2018.
[7] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112. .
[8] L. Yu, W. Zhang, J. Wang, Y. Yu, Seqgan: Sequence generative adversarial nets with policy gradient, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017. .
[9] J.D.M.-W.C. Kenton, L.K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[10] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf (2018). .
[11] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, M. Huang, Augmenting end-to-end dialogue systems with commonsense knowledge, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018. .
[12] L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao, W.B. Croft, X. Liu, Y. Shen, J. Liu, A hybrid retrieval-generation neural conversation model, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 1341–1350.
[13] S. Liu, H. Chen, Z. Ren, Y. Feng, Q. Liu, D. Yin, Knowledge diffusion for neural dialogue generation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1489–1498.
[14] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, X. Zhu, Commonsense knowledge aware conversation generation with graph attention., in: IJCAI, 2018, pp. 4623–4629.
[15] S. Moon, P. Shah, A. Kumar, R. Subba, Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 845–854.
[16] H. Zhang, Z. Liu, C. Xiong, Z. Liu, Grounded conversation generation as guided traverses in commonsense knowledge graphs, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2031–2043.
[17] J. Jung, B. Son, S. Lyu, Attnio: Knowledge graph exploration with in-and-out attention flow for knowledge-grounded dialogue, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 3484–3497.
[18] A. Madotto, C.-S. Wu, 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 (Volume 1: Long Papers), 2018, pp. 1468–1478.
[19] C.-S. Wu, R. Socher, C. Xiong, Global-to-local memory pointer networks for task-oriented dialogue, in: 7th International Conference on Learning Representations, ICLR 2019, 2019. .
[20] S. Wu, Y. Li, D. Zhang, Y. Zhou, Z. Wu, Diverse and informative dialogue generation with context-specific commonsense knowledge awareness, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5811–5820.
[21] S. Wu, Y. Li, D. Zhang, Z. Wu, Improving knowledge-aware dialogue response generation by using human-written prototype dialogues, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 1402–1411.
[22] S. Wu, Y. Li, D. Zhang, Y. Zhou, Z. Wu, Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI, 2020, pp. 3766–3772.
[23] S. Yang, R. Zhang, S. Erfani, Graphdialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1878–1888.
[24] L. Xu, Q. Zhou, K. Gong, X. Liang, J. Tang, L. Lin, End-to-end knowledge-routed relational dialogue system for automatic diagnosis, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 7346–7353. .
[25] J. Xu, H. Wang, Z. Niu, H. Wu, W. Che, Knowledge graph grounded goal planning for open-domain conversation generation., in: AAAI, 2020, pp. 9338–9345.
[26] J. Wang, J. Liu, W. Bi, X. Liu, K. He, R. Xu, M. Yang, Improving knowledge-aware dialogue generation via knowledge base question answering, in: AAAI, 2020, pp. 9169–9176.

[27] D. Raghu, N. Gupta, et al., Disentangling language and knowledge in task-oriented dialogs, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 1239–1255.
[28] A. Madotto, S. Cahyawijaya, G.I. Winata, Y. Xu, Z. Liu, Z. Lin, P. Fung, Learning knowledge bases with parameters for task-oriented dialogue systems, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 2372–2394.
[29] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Neural Information Processing Systems (NIPS), 2013, pp. 1–9.
[30] O. Vinyals, M. Fortunato, N. Jaitly, Pointer networks, in: Advances in Neural Information Processing Systems 28, 2015, pp. 2692–2700.
[31] R. Lian, M. Xie, F. Wang, J. Peng, H. Wu, Learning to select knowledge for response generation in dialog systems, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 5081–5087.
[32] X. Chen, F. Meng, P. Li, F. Chen, S. Xu, B. Xu, J. Zhou, Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 3426–3437.
[33] B. Kim, J. Ahn, G. Kim, Sequential latent knowledge selection for knowledge-grounded dialogue, in: International Conference on Learning Representations, 2019. .
[34] Y. Zhang, P. Ren, M. de Rijke, Improving background based conversation with context-aware knowledge pre-selection, arXiv preprint arXiv:1906.06685 (2019). .
[35] P. Ren, Z. Chen, C. Monz, J. Ma, M. de Rijke, Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 8697–8704. .
[36] C. Zheng, Y. Cao, D. Jiang, M. Huang, Difference-aware knowledge selection for knowledge-grounded conversation generation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 115–125.
[37] C. Meng, P. Ren, Z. Chen, W. Sun, Z. Ren, Z. Tu, M.d. Rijke, Dukenet: A dual knowledge interaction network for knowledge-grounded conversation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1151–1160.
[38] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, J. Weston, Wizard of wikipedia: Knowledge-powered conversational agents, in: International Conference on Learning Representations, 2018. .
[39] S. Yavuz, A. Rastogi, G.-L. Chao, D. Hakkani-Tur, Deepcopy: Grounded response generation with hierarchical pointer networks, in: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, 2019, pp. 122–132.
[40] C. Meng, P. Ren, Z. Chen, C. Monz, J. Ma, M. de Rijke, Refnet: A reference-aware network for background based conversation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 8496–8503. .
[41] X. Lin, W. Jian, J. He, T. Wang, W. Chu, Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 41–52.
[42] X. Zhao, C. Tao, W. Wu, C. Xu, D. Zhao, R. Yan, A document-grounded matching network for response selection in retrieval-based chatbots, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 5443–5449.
[43] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, J. Zhou, Incremental transformer with deliberation decoder for document grounded conversations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 12–21.
[44] X. Zhao, L. Wang, R. He, T. Yang, J. Chang, R. Wang, Multiple knowledge syncretic transformer for natural dialogue generation, in: Proceedings of The Web Conference 2020, 2020, pp. 752–762.
[45] L. Qin, M. Galley, C. Brockett, X. Liu, X. Gao, B. Dolan, Y. Choi, J. Gao, Conversing by reading: Contentful neural conversation with on-demand machine reading, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5427–5436.
[46] Z. Tian, W. Bi, D. Lee, L. Xue, Y. Song, X. Liu, N.L. Zhang, Response-anticipated memory for on-demand knowledge integration in response generation, arXiv preprint arXiv:2005.06128 (2020). .
[47] Y. Wang, W. Rong, J. Zhang, Y. Ouyang, Z. Xiong, Knowledge grounded pre-trained model for dialogue response generation, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.
[48] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, R. Yan, Knowledge-grounded dialogue generation with pre-trained language models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 3377–3390.
[49] Y.-L. Tuan, W. Wei, W.Y. Wang, Unsupervised injection of knowledge into dialogue generation via language models, arXiv preprint arXiv:2004.14614 (2020). .
[50] X. Zhao, W. Wu, C. Tao, C. Xu, D. Zhao, R. Yan, Low-resource knowledge-grounded dialogue generation, in: International Conference on Learning Representations, 2019. .
[51] J. Weston, S. Chopra, A. Bordes, Memory networks, arXiv preprint arXiv:1410.3916 (2014). .
[52] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2, 2015, pp. 2440–2448. .
[53] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). .
[54] K. Clark, M.-T. Luong, Q.V. Le, C.D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, in: International Conference on Learning Representations, 2019. .
[55] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, H.-W. Hon, Unified language model pre-training for natural language understanding and generation, in: Advances in Neural Information Processing Systems, 2019, pp. 13063–13075.
[56] N. Moghe, S. Arora, S. Banerjee, M.M. Khapra, Towards exploiting background knowledge for building conversation systems, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2322–2332.
[57] K. Zhou, S. Prabhumoye, A.W. Black, A dataset for document grounded conversations, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 708–713.
[58] Y.-L. Tuan, Y.-N. Chen, H.-Y. Lee, Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs, 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, pp. 1855–1865.
[59] W. Wu, Z. Guo, X. Zhou, H. Wu, X. Zhang, R. Lian, H. Wang, Proactive human-machine conversation with explicit conversation goal, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3794–3804.
[60] H. Zhou, C. Zheng, K. Huang, M. Huang, X. Zhu, Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7098–7108.
[61] H. Su, X. Shen, Z. Xiao, Z. Zhang, E. Chang, C. Zhang, C. Niu, J. Zhou, Moviechats: Chat like humans in a closed domain, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6605–6619.
[62] F. Jelinek, R.L. Mercer, L.R. Bahl, J.K. Baker, Perplexity–a measure of the difficulty of speech recognition tasks, The Journal of the Acoustical Society of America 62 (S1) (1977) S63.
[63] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[64] C.-Y. Lin, F. Och, Looking for a few good metrics: Rouge and its evaluation, in: Ntcir Workshop, 2004. .
[65] J. Li, M. Galley, C. Brockett, J. Gao, W.B. Dolan, A diversity-promoting objective function for neural conversation models, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 110–119.
[66] Z.-Y. Niu, H. Wu, H. Wang, et al., Knowledge aware conversation generation with explainable reasoning over augmented graphs, 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, pp. 1782–1792.
[67] D. Zhang, B. Guo, Z. Yu, The emergence of social and community intelligence, Computer 44 (7) (2011) 21–28.
[68] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020). .
[69] J. Guan, F. Huang, Z. Zhao, X. Zhu, M. Huang, A knowledge-enhanced pretraining model for commonsense story generation, Transactions of the Association for Computational Linguistics 8 (2020) 93–108.
[70] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, H. Wang, Ernie 2.0: A continual pre-training framework for language understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 8968–8975. .
[71] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, Ernie: Enhanced language representation with informative entities, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1441–1451.
[72] T. Shen, Y. Mao, P. He, G. Long, A. Trischler, W. Chen, Exploiting structured knowledge in text via graph-guided representation learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8980–8994.
[73] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning PMLR, 2017, pp. 1126–1135.
[74] K. Qian, Z. Yu, Domain adaptive dialog generation via meta learning, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2639–2649.
[75] F. Mi, M. Huang, J. Zhang, B. Faltings, Meta-learning for low-resource natural language generation in task-oriented dialogue systems, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI, 2019, pp. 3151–3157.
[76] S. Mazumder, N. Ma, B. Liu, Towards a continuous knowledge learning engine for chatbots, arXiv preprint arXiv:1802.06024 (2018). .

Hao Wang was born in 1996. He received his B.E. degree in computer science and technology from NWPU in 2019. He is currently working toward a Ph.D. degree at NWPU. His current research interests include natural language processing and dialog systems.

Bin Guo was born in 1980. He is a Ph.D. professor and Ph.D. supervisor. He is a senior member of China Computer Federation. His main research interests include ubiquitous computing, social and community intelligence, urban big data mining, mobile crowd sensing, and human computer interaction.

Wei Wu was born in 1985. He obtained his Ph.D. in Applied Mathematics from Peking University in 2012. His roles include principal applied scientist lead in Microsoft Xiaoice Team and lead researcher in Microsoft Research Asia (MSRA). His research interests include machine learning, natural language processing, and information retrieval.

Sicong Liu was born in 1992. She is currently an associate professor in the School of Computer Science at Northwestern Polytechnical University (NWPU). She is doing two-years postdoc at NWPU, supervised by Prof. Bin Guo on deep learning in AIoT, crowd computing for Human–Machine-IoT.

Zhiwen Yu was born in 1977. He is a Ph.D. professor and Ph.D. supervisor. He is a senior member of China Computer Federation. His main research interests include mobile internet, ubiquitous computing, social and community intelligence, urban big data mining, mobile crowd sensing, and human computer interaction.