Artificial Intelligence Review (2023) 56:3055–3155
https://doi.org/10.1007/s10462-022-10248-8

Recent advances in deep learning based dialogue systems: a systematic survey

Jinjie Ni[1] · **Tom Young**[1] · **Vlad Pandelea**[1] · **Fuzhao Xue**[1] · **Erik Cambria**[1]

Published online: 20 August 2022

Abstract
Dialogue systems are a popular natural language processing (NLP) task as it is promising in real-life applications. It is also a complicated task since many NLP tasks deserving study are involved. As a result, a multitude of novel works on this task are carried out, and most of them are deep learning based due to their outstanding performance. In this survey, we mainly focus on the deep learning based dialogue systems. We comprehensively review state-of-the-art research outcomes in dialogue systems and analyze them from two angles: model type and system type. Specifically, from the angle of model type, we discuss the principles, characteristics, and applications of different models that are widely used in dialogue systems. This will help researchers acquaint these models and see how they are applied in state-of-the-art frameworks, which is rather helpful when designing a new dialogue system. From the angle of system type, we discuss task-oriented and open-domain dialogue systems as two streams of research, providing insight into the hot topics related. Furthermore, we comprehensively review the evaluation methods and datasets for dialogue systems to pave the way for future research. Finally, some possible research trends are identified based on the recent research outcomes. To the best of our knowledge, this survey is the most comprehensive and up-to-date one at present for deep learning based dialogue systems, extensively covering the popular techniques. We speculate that this work is a good starting point for academics who are new to the dialogue systems or those who want to quickly grasp up-to-date techniques in this area.

Keywords Dialogue systems · Chatbots · Conversational AI · Natural language processing · Deep learning

Tom Young and Vlad Pandelea have equal contribution.

The frameworks, topics, and datasets discussed are originated from the extensive literature review of state-of-the-art research. We have tried our best to cover all but may still omit some works. Readers are welcome to provide suggestions regarding the omissions and mistakes in this article. We also intend to update this article with time as and when new approaches or definitions are proposed and used by the community.

✉ Erik Cambria
cambria@ntu.edu.sg

Extended author information available on the last page of the article

Table 1 Examples of inputs and outputs of task-oriented and open-domain dialogue systems in datasets

Category	User message (U)	Agent response (R)	External knowledge (K)
Task-oriented	I need to find a nice restaurant in Madrid that serves expensive Thai food	There is a restaurant called *Bangkok City* locating at 9 Red Ave.	Restaurant database
Open-domain	I love the grilled fish so much!	Yeah. it's a famous *Chinese dish*	Commonsense KG

Some datasets provide external knowledge annotations for each dialogue pair, e.g., in task-oriented dialogue systems, the external knowledge can be retrieved from restaurant databases; in open-domain dialogue systems, it can be retrieved from commonsense knowledge graphs (KG)

1 Introduction

Dialogue systems (or chatbots) are playing a bigger role in the world. People may still have a stereotype that chatbots are those rigid agents in their phone calls to a bank. However, thanks to the revival of artificial intelligence, the modern chatbots can converse with rich topics ranging from your birthday party to a speech given by Biden, and, if you want, they can even book a place for your party or play the speech video. At present, dialogue systems are one of the hot topics in NLP and are highly demanded in industry and daily life. The market size of chatbot is projected to grow from $2.6 billion in 2021 to $9.4 billion by 2024 at a compound annual growth rate (CAGR) of 29.7%[1] and 80% of businesses are expected to be equipped with chatbot automation by the end of 2021.[2]

Dialogue systems perform chit-chat with human or serve as an assistant via conversations. By their applications, dialogue systems are commonly divided into two categories: task-oriented dialogue systems (TOD) and open-domain dialogue systems (OOD). Task-oriented dialogue systems solve specific problems in a certain domain such as movie ticket booking, restaurant table reserving, etc. Instead of focusing on task completion, open-domain dialogue systems aim to chat with users without the task and domain restrictions (Ritter et al. 2011), which are usually fully data-driven. Both task-oriented and open-domain dialogue systems can be seen as a mapping φ from user message $U = \{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(i)}\}$ to agent response $R = \{\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \ldots, \mathbf{r}^{(j)}\}$: $R = \varphi(U)$, where $\mathbf{u}^{(i)}$ and $\mathbf{r}^{(j)}$ denote the ith token of the user message and the jth token of the agent response respectively. In many open-domain and task-oriented dialogue systems, this mapping also considers a source of external knowledge/database K as input: $R = \varphi(U, K)$. Table 1 presents examples of inputs and outputs of task-oriented and open-domain dialogue systems. More specific details and works will be discussed in Sects. 3 and 4.

Traditional task-oriented dialogue systems are organized in a pipeline structure and consist of four functional modules: Natural Language Understanding, Dialogue State Tracking, Policy Learning, and Natural Language Generation, which will be discussed in detail in Sect. 3. Many state-of-the-art works design end-to-end task-oriented dialogue systems to achieve better optimization compared with pipeline methods. Open-domain dialogue systems are generally divided into three categories: generative systems, retrieval-based systems, and ensemble systems. Generative systems apply sequence-to-sequence models (see

[1] Statistic source: https://markets.businessinsider.com.

[2] Statistic source: https://outgrow.co.

Sect. 2.2.5) to map the user message and dialogue history into a response sequence that may not appear in the training corpus. By contrast, retrieval-based systems try to select a pre-existing response from a certain response set. Ensemble systems combine generative methods and retrieval-based methods in two ways: retrieved responses can be compared with generated responses to choose the best among them; generative models can also be used to refine the retrieved responses (Zhu et al. 2019; Song et al. 2016; Qiu et al. 2017; Serban et al. 2017a). Generative systems can produce flexible and dialogue context-related responses while sometimes they lack coherence[3] and tend to make dull responses (Serban et al. 2016; Vinyals and Le 2015; Sordoni et al. 2015b). Retrieval-based systems select responses from human response sets and thus are able to achieve better coherence in surface-level language. However, retrieval systems are restricted by the finiteness of the response sets and sometimes the responses retrieved show a weak correlation with the dialogue context (Zhu et al. 2019).

For dialogue systems, existing surveys (Arora et al. 2013; Wang and Yuan 2016; Mallios and Bourbakis 2016; Chen et al. 2017a; Gao et al. 2018) are either outdated or not comprehensive. Some definitions in these papers are no longer being used at present, and a lot of new works and topics are not covered. In addition, most of them lack a multi-angle analysis. Thus, in this survey, we comprehensively review high-quality works in recent years with a focus on deep learning-based approaches and provide insight into state-of-the-art research from both model angle and system angle. Moreover, this survey updates the definitions/names according to state-of-the-art research. E.g., we name "open-domain dialogue systems" instead of "chit-chat dialogue systems" because most of the articles (roughly 70% according to our survey) name them as the prior one. We also extensively cover the diverse hot topics in dialogue systems and extend some new topics that are popular in current research community (such as Domain Adaptation, Dialogue State Tracking Efficiency, End-to-end methods for task-oriented dialogue systems; Controllable Generation, Interactive Training, and Visual Dialogue for open-domain dialogue systems).

Traditional dialogue systems are mostly rule-based (Arora et al. 2013) and non-neural machine learning based systems. Rule-based systems are easy to implement and can respond naturally, which contributed to their popularity in earlier industry products. However, the dialogue flows of these systems are predetermined, which keeps the applications of the dialogue systems within certain scenarios. Non-neural machine learning based systems usually perform template filling to manage certain tasks. These systems are more flexible compared with rule-based systems because the dialogue flows are not predetermined. However, they cannot achieve high F1 scores (Powers 2020) in template filling[4] and are also restricted in application scenarios and response diversity because of the fixed templates. Most if not all state-of-the-art dialogue systems are deep learning-based systems (neural systems). The rapid growth of deep learning improves the performance of dialogue systems (Chen et al. 2017a). Deep learning can be viewed as representation learning with multilayer neural networks. Deep learning architectures are widely used in dialogue systems and their subtasks. Section 2 discusses various popular deep learning architectures.

Apart from dialogue systems, there are also many dialogue-related tasks in NLP, including but not limited to question answering, reading comprehension, dialogue

[3] The quality of being logical and consistent not only between words/subwords but also between responses of different timesteps.

[4] Template filling is an efficient approach to extract and structure complex information from text to fill in a pre-defined template. They are mostly used in task-oriented dialogue systems.

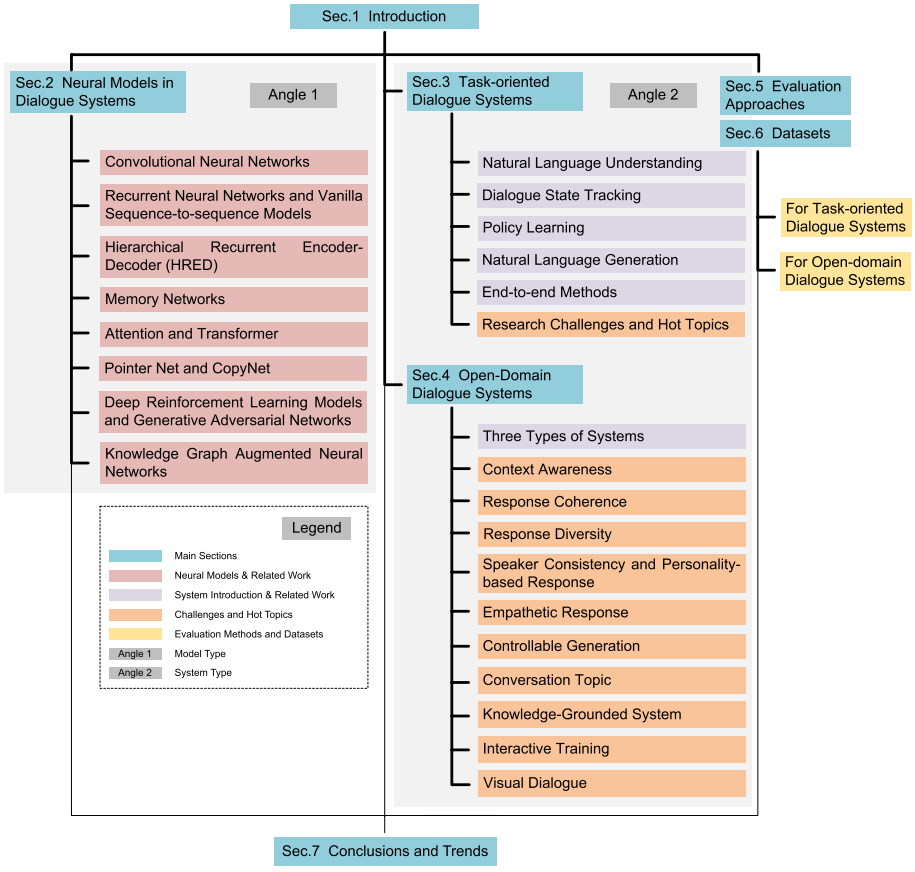


Fig. 1 The overall diagram of this article

disentanglement, visual dialogue, visual question answering, dialogue reasoning, conversational semantic parsing, dialogue relation extraction, dialogue sentiment analysis, hate speech detection, MISC detection, etc. In this survey, we also touch on some works tackling these dialogue-related tasks, since the design of dialogue systems can benefit from advances in these related areas.

We produced a diagram for this article to help readers familiarize the overall structure (Fig. 1). In this survey, Sect. 1 briefly introduces dialogue systems and deep learning; Sect. 2 discusses the neural models popular in modern dialogue systems and the related work; Sect. 3 introduces the principles and related work of task-oriented dialogue systems and discusses the research challenges and hot topics; Sect. 4 briefly introduces the three kinds of systems and then focuses on hot topics in open-domain dialogue systems; Sect. 5 reviews the main evaluation methods for dialogue systems; Sect. 6 comprehensively summarizes the datasets commonly used for dialogue systems; finally, Sect. 7 concludes the paper and provides some insight on research trends.

2 Neural models in dialogue systems

In this section, we introduce neural models that are popular in state-of-the-art dialogue systems and related subtasks. We also discuss the applications of these models or their variants in modern dialogue systems research to provide readers with a picture from the model's perspective. This will help researchers acquaint these models and see how they are applied in state-of-the-art frameworks, which is rather helpful when designing a new dialogue system. The models discussed include: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Vanilla Sequence-to-sequence Models, Hierarchical Recurrent Encoder-Decoder (HRED), Memory Networks, Attention Networks, Transformer, Pointer Net and CopyNet, Deep Reinforcement Learning models, Generative Adversarial Networks (GANs), Knowledge Graph Augmented Neural Networks. We start from some classical models (e.g., CNNs and RNNs), and readers who are familiar with their principles and corresponding applications in dialogue systems can choose to read selectively.

2.1 Convolutional neural networks

Deep neural networks have been considered as one of the most powerful models. 'Deep' refers to the fact that they are multilayer, which extracts features by stacking feed-forward layers. Feed-forward layers can be defined as: $y = \sigma(Wx + b)$. Where the σ is an activation function; W and b are trainable parameters. The feed-forward layers are powerful due to the activation function, which makes the otherwise linear operation, non-linear. Whereas there exist some problems when using feed-forward layers. Firstly, the operations of feed-forward layers or multilayer neural networks are just template matching, where they do not consider the specific structure of data. Furthermore, the fully connected mechanism of traditional multilayer neural networks causes an explosion in the number of parameters and thus leads to generalization problems. LeCun et al. (1998) proposed LeNet-5, an early CNN. The invention of CNNs mitigates the above problems to some extent.

CNNs (Fig. 2) usually consist of convolutional layers, pooling layers, and feed-forward layers. Convolutional layers apply convolution kernels to perform the convolution operation:

$$G(m,n) = (f * h)(m,n) = \sum_j \sum_k h(j,k) f(m-j, n-k) \tag{1}$$

where m and n are respectively the indexes of rows and columns of the result matrix. f denotes the input matrix and h denotes the convolutional kernel. The pooling layers perform down-sampling on the result of convolutional layers to get a higher level of features and the feed-forward layers map them into a probability distribution to predict class scores.

A sliding window feature enables convolution layers to capture local features and the pooling layers can produce hierarchical features. These two mechanisms give CNNs the local perception and global perception ability, helping to capture some specific inner structures of data. The parameter sharing mechanism eases the parameter explosion problem and overfitting problem because the reduction of trainable parameters leads to less model complexity, improving the generalization ability.

Due to these good properties, CNNs have been widely applied in many works. Among them, the Computer Vision tasks benefit the most for that the Spatio-temporal data structures of images or videos are perfectly captured by CNNs. For more detailed mechanism

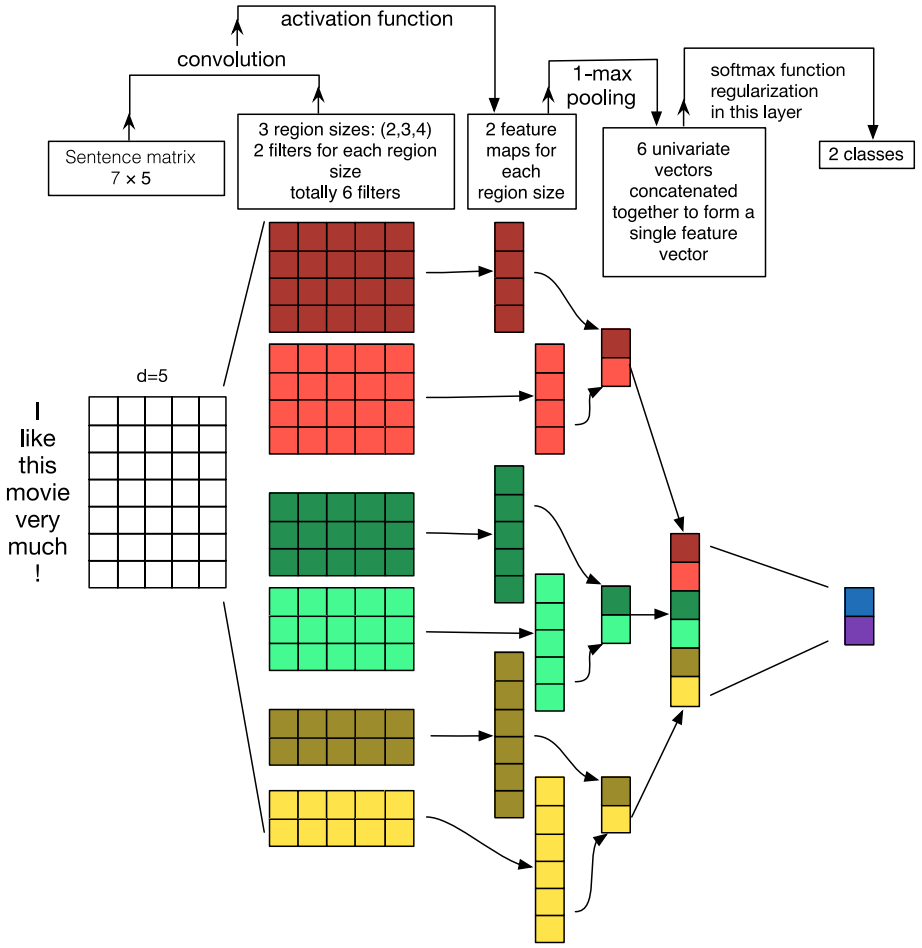


Fig. 2 A CNN architecture for text classification (Zhang and Wallace 2017)

illustrations and other variants of CNNs, readers can refer to these representative algorithm papers or surveys: (Krizhevsky et al. 2012; Zeiler and Fergus 2014; Simonyan and Zisserman 2015; Szegedy et al. 2015; He et al. 2016; Aloysius and Geetha 2017; Rawat and Wang 2017). In this survey, we focus on dialogue systems.

Recent years have seen a dramatic increase in applications of CNNs in NLP. Many tasks take words as basic units. However, phrases, sentences, or even paragraphs are also useful for semantic representations. As a result, CNNs are an ideal tool for the hierarchical modeling of language (Conneau et al. 2017).

CNNs are good textual feature extractors, but they may not be ideal sequential encoders. Some dialogue systems (Qiu et al. 2019; Bi et al. 2019; Ma et al. 2020a) directly used CNNs as the encoder of utterances or knowledge, but most of the state-of-the-art dialogue systems such as Feng et al. (2019), Wu et al. (2017), Tao et al. (2019), Wang et al. (2019b), Chauhan et al. (2019), Feldman and El-Yaniv (2019), Chen et al. (2019b), Lu et al. (2019b) and Coope et al. (2020) chose to use CNNs as a hierarchical feature extractor after encoding the text information, instead of directly applying them as encoders. This is due to the

fixed input length and limited convolution span of CNNs. Generally, there are two main situations where CNNs are used to process encoded information in dialogue systems. The first situation is applying CNNs to extract features directly based on the feature vectors from the encoder (Wang et al. 2019b; Chauhan et al. 2019; Feldman and El-Yaniv 2019; Chen et al. 2019b) and Coope et al. (2020). Within the works above, Feldman and El-Yaniv (2019) extracted features from character-level embeddings, illustrating the hierarchical extraction capability of CNNs. Another situation in which CNNs are used is extracting feature maps in response retrieval tasks. Some works built retrieval-based dialogue systems (Wu et al. 2017; Feng et al. 2019; Tao et al. 2019; Lu et al. 2019b). They used separate encoders to encode dialogue context and candidate responses and then used a CNN as an extractor of the similarity matrix calculated from the encoded dialogue context and candidate responses. Their experiments showed that this method can achieve good performance in response retrieval tasks.

The main reason why more recent works do not choose CNNs as dialogue encoders is that they fail to extract the information across temporal sequence steps continuously and flexibly (Krizhevsky et al. 2012). Some models introduced later do not process data points independently, which are desirable models for encoders.

2.2 Recurrent neural networks and vanilla sequence-to-sequence models

NLP tasks including dialogue-related tasks try to process and analyze sequential language data points. Even though standard neural networks, as well as CNNs, are powerful learning models, they have two main limitations (Lipton et al. 2015). One is that they assume the data points are independent of each other. While it is reasonable if the data points are produced independently, essential information can be missed when processing interrelated data points (e.g., text, audio, video). Additionally, their inputs are usually of fixed length, which is a limitation when processing sequential data varying in length. Thus, a sequential model being able to represent the sequential information flow is desirable.

Markov models like Hidden Markov Models (HMMs) are traditional sequential models, but due to the time complexity of the inference algorithm (Viterbi 1967) and because the size of transition matrix grows significantly with the increase of the discrete state space, in practice they are not applicable in dealing with problems involving large possible hidden states. The property that the hidden states of Markov models are only affected by the immediate hidden states further limits the power of this model.

RNN models are not proposed recently, but they greatly solve the above problems and some variants can amazingly achieve state-of-the-art performance in dialogue-related tasks as well as many other NLP tasks. The inductive bias of recurrent models is non-replaceable in many scenarios, and many up-to-date models incorporate the recurrence.

2.2.1 Jordan-type and Elman-type RNNs

In 1982, Hopfield introduced an early family of RNNs to solve pattern recognition tasks (Hopfield 1982). Jordan (1986) and Elman (1990) introduced two kinds of RNN architectures respectively. Generally, modern RNNs can be classified into Jordan-type RNNs and Elman-type RNNs.

The Jordan-type RNNs are shown in Fig. 3a. x_t, h_t, and y_t are the inputs, hidden state, and output of time step t, respectively. W_h, W_y, and U_h are weight matrices. Each update of the hidden state is decided by the current input and the output of the last time step

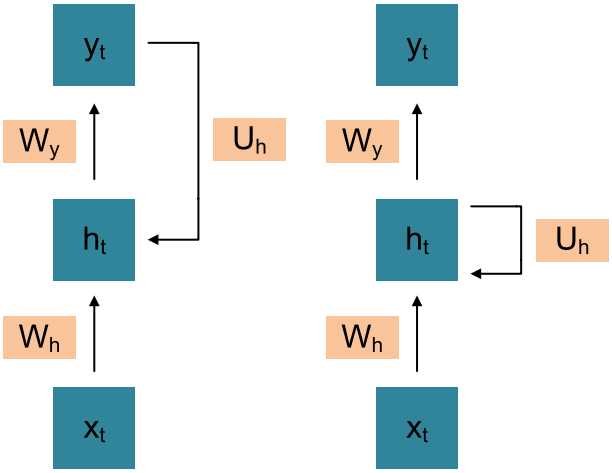


Fig. 3 Graphical models of two basic types of RNNs

(a) Jordan-type RNNs **(b)** Elman-type RNNs

while each output is decided by the current hidden state. Thus the hidden state and output of time step t can be calculated as:

$$h_t = \sigma_h\left(W_h x_t + U_h y_{t-1} + b_h\right) \tag{2}$$

$$y_t = \sigma_y\left(W_y h_t + b_y\right) \tag{3}$$

where b_h and b_y are biases. σ_h and σ_y are activation functions.

The Elman-type RNNs are shown in Fig. 3b. The difference is that each hidden state is decided by the current input and the hidden state of last time step. Thus the hidden state and output of time step t can be calculated as:

$$h_t = \sigma_h\left(W_h x_t + U_h h_{t-1} + b_h\right) \tag{4}$$

$$y_t = \sigma_y\left(W_y h_t + b_y\right) \tag{5}$$

Simple RNNs can model long-term dependencies theoretically. But in practical training, long-range dependencies are difficult to learn (Bengio et al. 1994; Hochreiter et al. 2001). When backpropagating errors over many time steps, simple RNNs suffer from problems known as gradient vanishing and gradient explosion (Hochreiter and Schmidhuber 1997). Some solutions were proposed to solve these problems (Williams and Zipser 1989; Pascanu et al. 2013), which led to the inventions of some variants of traditional recurrent networks.

2.2.2 LSTM

Hochreiter and Schmidhuber (1997) introduced gate mechanisms in LSTM mainly to address the gradient vanishing problem. Input gate, forget gate and output gate were introduced to decide how much information from new inputs and past memories should be reserved. The model can be described by the following equations:

$$\hat{h}^{(t)} = \tanh\left(W^{\hat{h}x}x^{(t)} + W^{\hat{h}h}h^{(t-1)} + b_{\hat{h}}\right) \tag{6}$$

$$i^{(t)} = \sigma\left(W^{ix}x^{(t)} + W^{ih}h^{(t-1)} + b_i\right) \tag{7}$$

$$f^{(t)} = \sigma\left(W^{fx}x^{(t)} + W^{fh}h^{(t-1)} + b_f\right) \tag{8}$$

$$o^{(t)} = \sigma\left(W^{ox}x^{(t)} + W^{oh}h^{(t-1)} + b_o\right) \tag{9}$$

$$s^{(t)} = \hat{h}^{(t)} \odot i^{(t)} + s^{(t-1)} \odot f^{(t)} \tag{10}$$

$$h^{(t)} = \tanh\left(s^{(t)}\right) \odot o^{(t)} \tag{11}$$

where t represents time step t. i, f and o are gates, denoting input gate, forget gate and output gate respectively. x, $\hat{h}$, s and h are input, short-term memory, long-term memory and output respectively. b is bias and W is weight matrix. $\odot$ denotes element-wise multiplication.

The intuition of the term "Long Short-Term Memory" is that the proposed model applies both long-term and short-term memory vectors to encode the sequential data, and uses gate mechanisms to control the information flow. The performance of LSTM is impressive since that it achieved state-of-the-art results in many NLP tasks as a backbone model although this model was proposed in 1997.

2.2.3 GRU

Inspired by the gating mechanism, Cho et al. (2014b) proposed Gated Recurrent Unit (GRU), which can be modeled by the equations:

$$z^{(t)} = \sigma\left(W^z x^{(t)} + U^z h^{(t-1)} + b_z\right) \tag{12}$$

$$r^{(t)} = \sigma\left(W^r x^{(t)} + U^r h^{(t-1)} + b_r\right) \tag{13}$$

$$\hat{h}^{(t)} = tanh\left(W^h x^{(t)} + U^h(r^{(t)} \odot h^{(t-1)}) + b_h\right) \tag{14}$$

$$h^{(t)} = (1 - z^{(t)}) \odot h^{(t-1)} + z^{(t)} \odot \hat{h}^{(t)} \tag{15}$$

where t represents time step t. z and r are gates, denoting update gate and reset gate respectively. x, $\hat{h}$ and h are input, candidate activation vector and output respectively. b is bias while W and U are weight matrixes. $\odot$ denotes element-wise multiplication.

LSTM and GRU, as two types of gating units, are very similar to each other (Chung et al. 2014). The most prominent common point between them is that from time step t to time step $t + 1$, an additive component is introduced to update the state whereas simple RNNs always replace the activation. Both LSTM and GRU keep certain old components and mix them with new contents. This property enables the units to remember the

information of history steps farther back and, more importantly, avoid gradient vanishing problems when backpropagating the error.

There also exist several differences between them. LSTM exposes its memory content under the control of the output gate, while the same content in GRU is in an uncontrolled manner. Additionally, different from LSTM, GRU does not independently gate the amount of new memory content being added. And if looking from experimental perspective, GRU has fewer parameters, which contributes to its faster convergence and better generalization ability. It has also been shown that GRU can achieve better performance in smaller datasets (Chung et al. 2014). However, Gruber and Jockisch (2020) showed that LSTM cells exhibited consistently better performance in a large-scale analysis of Neural Machine Translation.

2.2.4 Bidirectional recurrent neural networks

In sequence learning, not only the past information is essential to the model inference, but the future information should also be considered to achieve a better inference ability. Schuster and Paliwal (1997) proposed the bi-directional recurrent neural networks (BRNNs), which had two kinds of hidden layers: the first encoded information from past time steps while the second encoded information in a flipped direction. The model can be described using the equations:

$$h^{(t)} = \sigma\left(W^{hx}x^{(t)} + W^{hh}h^{(t-1)} + b_h\right) \tag{16}$$

$$z^{(t)} = \sigma\left(W^{zx}x^{(t)} + W^{zz}z^{(t+1)} + b_z\right) \tag{17}$$

$$\hat{y}^{(t)} = softmax\left(W^{yh}h^{(t)} + W^{yz}z^{(t)} + b_y\right) \tag{18}$$

where h and z are the two hidden layers. Other variables are defined in the same way as in the case of LSTMs and GRUs.

2.2.5 Vanilla sequence-to-sequence models (encoder–decoder models)

Sutskever et al. (2014) first proposed the sequence-to-sequence model to solve the machine translation tasks. The sequence-to-sequence model aimed to map an input sequence to an output sequence by first using an encoder to map the input sequence into an intermediate vector and a decoder further generated the output based on the intermediate vector and history generated by the decoder. The equations below illustrate the encoder-decoder model:

$$Encoder : h_t = E(h_{t-1}, x_t) \tag{19}$$

$$Decoder : y_t = D(h_t, y_{t-1}) \tag{20}$$

wt is the time step, h is the hidden vector and y is the output vector. E and D are the sequential cells used by the encoder and decoder respectively. The last hidden state of the encoder is the intermediate vector, and this vector is usually used to initialize the first hidden state of the decoder. At encoding time, each hidden state is decided by the hidden state of the previous time step and the input at the current time step, while at decoding time, each hidden state is decided by the current hidden state and the output of the previous time step.

This model is powerful because it is not restricted to fixed-length inputs and outputs. Instead, the length of the source sequence and target sequence can differ. Based on this model, many more advanced sequence-to-sequence models have been developed, which will be discussed in this and subsequent sections.

RNNs play an essential role in neural dialogue systems for their strong ability to encode sequential text information. RNNs and their variants are found in many dialogue systems. Task-oriented systems apply RNNs as encoders of dialogue context, dialogue state, knowledge base entries, and domain tags (Moon et al. 2019; Chen et al. 2019a; Wu et al. 2019b, a). Open-domain systems apply RNNs as dialogue history encoders (Sankar et al. 2019; Du and Black 2019; Ji et al. 2020; Chen et al. 2020c), among which retrieval-based systems model dialogue history and candidate responses together (Zhu et al. 2019; Tang et al. 2019; Feldman and El-Yaniv 2019; Lu et al. 2019b). In knowledge-grounded systems, RNNs are encoders of outside knowledge sources (e.g., background, persona, topic, etc.) (Shuster et al. 2020c; Majumder et al. 2020b; Chen et al. 2020c; Cho and May 2020).

Furthermore, as the decoder of sequence-to-sequence models in dialogue systems (Huang et al. 2020c; Song et al. 2019; Liu et al. 2019; Lin et al. 2019), RNNs usually decode the hidden state of utterance sequences by greedy search or beam search (Aubert et al. 1994). These decoding mechanisms cause problems like generic responses, which will be discussed in later sections.

Some works (Liu et al. 2019; Mehri et al. 2019; Chen et al. 2019b; Ma et al. 2020a) combined RNNs as a part of dialogue representation models to train dialogue embeddings and further improved the performance of dialogue-related tasks. These embedding models were trained on dialogue tasks and present more dialogue features. They consistently outperformed state-of-the-art contextual representation models (e.g., BERT, ELMo, and GPT) in some dialogue tasks when these contextual representation models were not fine-tuned for the specific tasks.

2.3 Hierarchical recurrent encoder–decoder (HRED)

Hierarchical Recurrent Encoder–Decoder (HRED) is a context-aware sequence-to-sequence model. It was first proposed by Sordoni et al. (2015a) to address the context-aware online query suggestion problem. It was designed to be aware of historical queries and the proposed model can provide rare and high-quality results.

With the popularity of the sequence-to-sequence model, Serban et al. (2016) extended HRED to the dialogue domain and built an end-to-end context-aware dialogue system. HRED achieved noticeable improvements in dialogue and end-to-end question answering. This work attracted even more attention than the original paper for that dialogue systems are a perfect setting for the application of HRED. Traditional dialogue systems (Ritter et al. 2011) generated responses based on the single-turn messages, which sacrificed the information in the dialogue history. Sordoni et al. (2015b) combined dialogue history turns with a window size of 3 as the input of a sequence-to-sequence model for response generation, which is limited as well for that they encode the dialogue history only in token-level. The “turn-by-turn" characteristic of dialogue indicated that the turn-level information also matters. The HRED learned both token-level and turn-level representation, thus exhibiting promising dialogue context awareness.

Figure 4 represents the HRED in a dialogue setting. HRED models the token-level and turn-level sequences hierarchically with two levels of RNNs: a token-level RNN consisting of an encoder and a decoder, and a turn-level context RNN. The encoder RNN encodes the

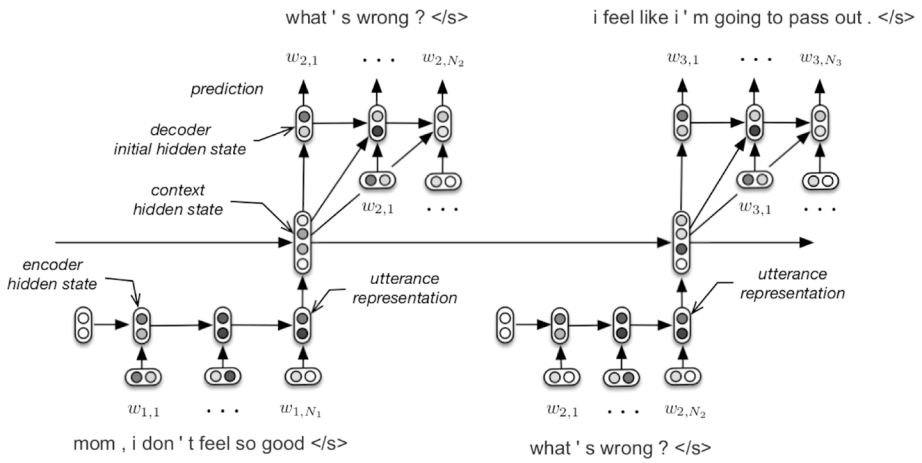


Fig. 4 The HRED model in a dialogue setting (Serban et al. 2016)

utterance of each turn token by token into a hidden state. This hidden state is then taken as the input of the context RNN at each turn-level time step. Thus the turn-level context RNN iteratively keeps track of the history utterances. The hidden state of context RNN at turn t represents a summary of the utterances up to turn t and is used to initialize the first hidden state of decoder RNN, which is similar to a standard decoder in sequence-to-sequence models (Sutskever et al. 2014). All of the three RNNs described above apply GRU cells as the recurrent unit, and the parameters of encoder and decoder are shared for each utterance.

Serban et al. (2017b) further proposed Latent Variable Hierarchical Recurrent Encoder–Decoder (VHRED) to model complex dependencies between sequences. Based on HRED, VHRED combined a latent variable into the decoder and turned the decoding process into a two-step generation process: sampling a latent variable at the first step and then generating the response conditionally. VHRED was trained with a variational lower bound on the log-likelihood and exhibited promising improvement in diversity, length, and quality of generated responses.

Many recent works in dialogue-related tasks apply HRED-based frameworks to capture hierarchical dialogue features. Zhang et al. (2019c) argued that standard HRED processed all contexts in dialogue history indiscriminately. Inspired by the architecture of Transformer (Vaswani et al. 2017), they proposed ReCoSa, a self-attention-based hierarchical model. It first applied LSTM to encode token-level information into context hidden vectors and then calculated the self-attention for both the context vectors and masked response vectors. At the decoding stage, the encoder–decoder attention was calculated to facilitate the decoding. Shen et al. (2019) proposed a hierarchical model consisting of 3 hierarchies: the discourse-level which captures the global knowledge, the pair-level which captured the topic information in utterance pairs, and the utterance level which captured the content information. Such a multi-hierarchy structure contributed to its higher quality responses in terms of diversity, coherence, and fluency. Chauhan et al. (2019) applied HRED and VGG-19 as a multimodal HRED (MHRED). The HRED encoded hierarchical dialogue context while VGG-19 extracted visual features for all images in the corresponding turn. With the addition of a position-aware attention mechanism, the model showed more diverse and accurate responses in a visually grounded setting. Mehri et al. (2019) learned

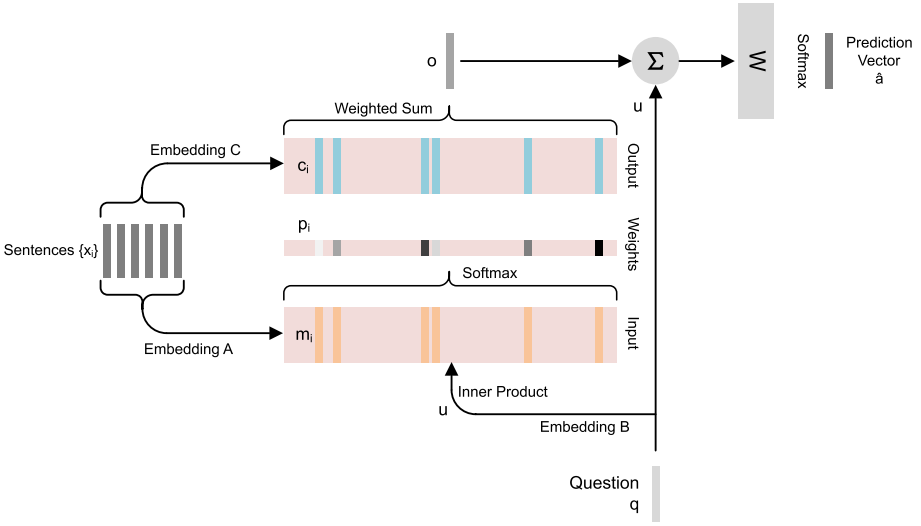


Fig. 5 The structure of end-to-end memory networks (Sukhbaatar et al. 2015)

dialogue context representations via four sub-tasks, three of which (next-utterance generation, masked-utterance retrieval, and inconsistency identification) made uses of HRED as the context encoder, and good performance was achieved. Cao et al. (2019) used HRED to encode the dialogue history between therapists and patients to categorize therapist and client MI behavioral codes and predict future codes. Qiu et al. (2020) applied an LSTM-based VHRED to address the two-agent and multi-agent dialogue structure induction problem in an unsupervised fashion. On top of that, they applied a Conditional Random Field model in two-agent dialogues and a non-projective dependency tree in multi-agent dialogues, both of them achieving better performance in dialogue structure modeling.

2.4 Memory networks

Memory is a crucial component when addressing problems regarding past experiences or outside knowledge sources. The hippocampus of human brains and the hard disk of computers are the components that humans and computers depend on for reading and writing memories. Traditional models rarely have a memory component, thus lacking the ability of knowledge reusing and reasoning. RNNs iteratively pass history information across time steps, which, to some extent, can be viewed as a memory model. However, even for LSTM, which is a powerful variant of RNN equipped with a long-term and short-term memory, the memory module is too small and facts are not explicitly discriminated, thus not being able to compress specific knowledge facts and reuse them in tasks.

Weston et al. (2015) proposed memory networks, a model that is endowed with a memory component. As described in their work, a memory network has five modules: a memory module which stores the representations of memory facts; an 'I' module which maps the input memory facts into embedded representations; a 'G' module which decides the update of the memory module; an 'O' module which generates the output conditioned on the input representation and memory representation; an 'R' module which organizes the

final response based on the output of 'O' module. This model needs a strong supervision signal for each module and thus is not practical to train in an end-to-end fashion.

Sukhbaatar et al. (2015) extended their prior work to an end-to-end memory network, which was commonly accepted as a standard memory network being easy to train and apply.

Figure 5 represents the proposed end-to-end memory networks. Its architecture consists of three stages: weight calculation, memory selection, and final prediction.

2.4.1 Weight calculation

The model first converts the input memory set $\{x_i\}$ into memory representations $\{m_i\}$ using a representation model A. Then it maps the input query into its embedding space using another representation model B, obtaining an embedding vector u. The final weights are calculated as follows:

$$p_i = Softmax(u^T m_i) \tag{21}$$

where p_i is the weight corresponding to each input memory x_i conditioned on the query.

2.4.2 Memory selection

Before generating the final prediction, a selected memory vector is generated by first encoding the input memory x_i into an embedded vector c_i using another representation model C, then calculating the weighted sum over the $\{c_i\}$ using the weights calculated in the previous stage:

$$o = \sum_i p_i c_i \tag{22}$$

where o represents the selected memory vector. This vector cannot be found in memory representations. The soft memory selection facilitates differentiability in gradient computing, which makes the whole model end-to-end trainable.

2.4.3 Final prediction

The final prediction is obtained by mapping the sum vector of the selected memory o and the embedded query u into a probability vector $\hat{a}$:

$$\hat{\alpha} = Softmax(W(o + u)) \tag{23}$$

Many dialogue-related works incorporate memory networks into their framework, especially for tasks involving an external knowledge base like task-oriented dialogue systems, knowledge-grounded dialogue systems, and QA.

Memory networks for task-oriented dialogue systems Chen et al. (2019b) argued that state-of-the-art task-oriented dialogue systems tended to combine dialogue history and knowledge base entries in a single memory module, which influenced the response quality. They proposed a task-oriented system that consists of three memory modules: two long-term memory modules storing the dialogue history and the knowledge base respectively; a working memory module that memorizes two distributions and controls the final word prediction. He et al. (2020a) trained a task-oriented dialogue system with a

"Two-teacher-one-student" framework to improve the knowledge retrieval and response quality of their memory networks. They first trained two teacher networks using reinforcement learning with complementary goal-specific reward functions respectively. Then with a GAN framework, they trained two discriminators to teach the student memory network to generate responses similar to those of the teachers, transferring the expert knowledge from the two teachers to the student. The advantage is that this training framework needs only weak supervision and the student network can benefit from the complementary targets of teacher networks. Kim et al. (2020b) solved the dialogue state tracking in task-oriented dialogue systems with a memory network that memorized the dialogue states. Different from other works, they did not update all dialogue states in the memory module from scratch. Instead, their model first predicted which states needed to be updated and then overwrote the target states. By selectively overwriting the memory module, they improved the efficiency of the dialogue state tracking task. Dai et al. (2020) applied the MemN2N (Sukhbaatar et al. 2015) as task-oriented utterance encoder, memorizing the existing responses and dialogue history. Then they used model-agnostic meta-learning (MAML) (Finn et al. 2017) to train the framework to retrieve correct responses in a few-shot fashion.

Memory networks for open-domain dialogue systems Tian et al. (2019) proposed a knowledge-grounded chit-chat system. A memory network was used to store query-response pairs and at the response generation stage, the generator produced the response conditioned on both the input query and memory pairs. It extracted key-value information from the query-response pairs in memory and combined them into token prediction. Xu et al. (2019) proposed to use meta-words to generate responses in open-domain systems in a controllable way. Meta-words are phrases describing response attributes. Using a goal-tracking memory network, they memorized the meta-words and generated responses based on the user message while incorporating meta-words at the same time. Gan et al. (2019) performed multi-step reasoning conditioned on a dialogue history memory module and a visual memory module. Two memory modules recurrently refined the representation to perform the next reasoning process. Experimental results illustrated the benefits of combining image and dialogue clues to improve the performance of visual dialogue systems. Han et al. (2019) trained a reinforcement learning agent to decide which memory vector can be replaced when the memory module is full to improve the accuracy and efficiency of the document-grounded question-answering task. They solved the scalability problem of memory networks by learning the query-specific value corresponding to each memory. Gao et al. (2020c) solved the same problem in a conversational machine reading task. They proposed an Explicit Memory Tracker (EMT) to decide whether the provided information in memory is enough for final prediction. Furthermore, a coarse-to-fine strategy was applied for the agent to make clarification questions to request additional information and refine the reasoning.

2.5 Attention and transformer

As introduced in Sect. 2.2, traditional sequence-to-sequence models decode the token conditioning on the current hidden state and output vector of last time step, which is formulated as:

$$P(y_i|y_1, \ldots, y_{i-1}, x) = g(y_{i-1}, h_i) \tag{24}$$

where g is a sequential model which maps the input vectors into a probability vector.

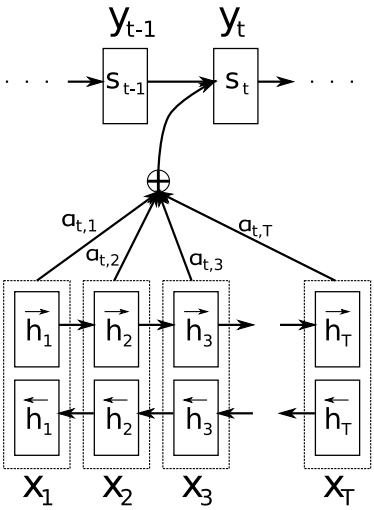


Fig. 6 The attention model (Bahdanau et al. 2015)

However, such a decoding scheme is limited when the input sentence is long. RNNs are not able to encode all information into a fixed-length hidden vector. Cho et al. (2014a) proved via experiments that a sequence-to-sequence model performed worse when the input sequence got longer. Also, for the limited-expression ability of a fixed-length hidden vector, the performance of the decoding scheme in Equation (24) largely depends on the first few steps of decoding, and if the decoder fails to have a good start, the whole sequence would be negatively affected.

2.5.1 Attention

Bahdanau et al. (2015) proposed the attention mechanism in the machine translation task. They described the method as "jointly align and translate", which illustrated the sequence-to-sequence translation model as an encoder-decoder model with attention. At the decoding stage, each decoding state would consider which parts of the encoded source sentence are correlated, instead of depending only on the immediate prior output token. The output probability distribution can be described as:

$$P(y_i|y_1, \ldots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i) \tag{25}$$

where i denotes the ith time step; y_i is the output token, s_i is the decoder hidden state and c_i is the weighted source sentence:

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \tag{26}$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{27}$$

Where α_{ij} is the normalized weight score:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{28}$$

e_{ij} is the similarity score between s_{i-1} and j^{th} encoder hidden state h_j, where the score is predicted by the similarity model a:

$$e_{ij} = a(s_{i-1}, h_j) \tag{29}$$

Figure 6 illustrates the attention model, where t and T denote time steps of decoder and encoder respectively.

Memory networks are similar to attention networks in the way they operate, except for the choice of the similarity model. In memory networks, the encoded memory can be viewed as the encoded source sentence in attention. However, the memory model proposed by Sukhbaatar et al. (2015) chose cosine distance as the similarity model while the attention proposed by Bahdanau et al. (2015) used a feed-forward network which is trainable together with the whole sequence-to-sequence model.

2.5.2 Transformer

Before transformers, most works combined attention with recurrent units, except for few works such as Parikh et al. (2016) and Gehring et al. (2017). Recurrent models condition each hidden state on the previous hidden state and the current input and are flexible in sequence length. However, due to their sequential nature, recurrent models cannot be trained in parallel, which severely undermines their potential. Vaswani et al. (2017) proposed Transformer, which entirely utilized attention mechanisms without any recurrent units and deployed more parallelization to speed up training. It applied self-attention and encoder-decoder attention to achieve local and global dependencies respectively.

Figure 7 represents the transformer. The following details its key mechanisms.

Encoder–decoder The Transformer consists of an encoder and a decoder. The encoder maps the input sequence $(x_1, \ldots, x_n)$ into continuous hidden states $(z_1, \ldots, z_n)$. The decoder further generates the output sequence $(y_1, \ldots, y_n)$ based on the hidden states of the encoder. The probability model of the Transformer is in the same form as that of the vanilla sequence-to-sequence model introduced in Sect. 2.2.5. Vaswani et al. (2017) stacked 6 identical encoder layers and 6 identical decoder layers. An encoder layer consists of a multi-head attention component and a simple feed-forward network, both of which apply residual structure. The structure of a decoder layer is almost the same as that of an encoder layer, except for an additional encoder-decoder attention layer, which computes the attention between decoder hidden states of the current time step and the encoder output vectors. The input of the decoder is partially masked to make sure that each prediction is based on the previous tokens, avoiding predicting with the presence of future information. Both inputs of encoder and decoder use a positional encoding mechanism.

Self-attention For an input sentence $x = (x_1, \ldots, x_n)$, each token x_i corresponds to three vectors: query, key, and value. The self-attention computes the attention weight for every token x_i against all other tokens in x by multiplying the query of x_i with the keys of all the remaining tokens one by one. For parallel computing, the query, key, and value vectors of all tokens are combined into three matrices: Query (Q), Key (K), and Value (V). The self-attention of an input sentence x is computed by the following formula:

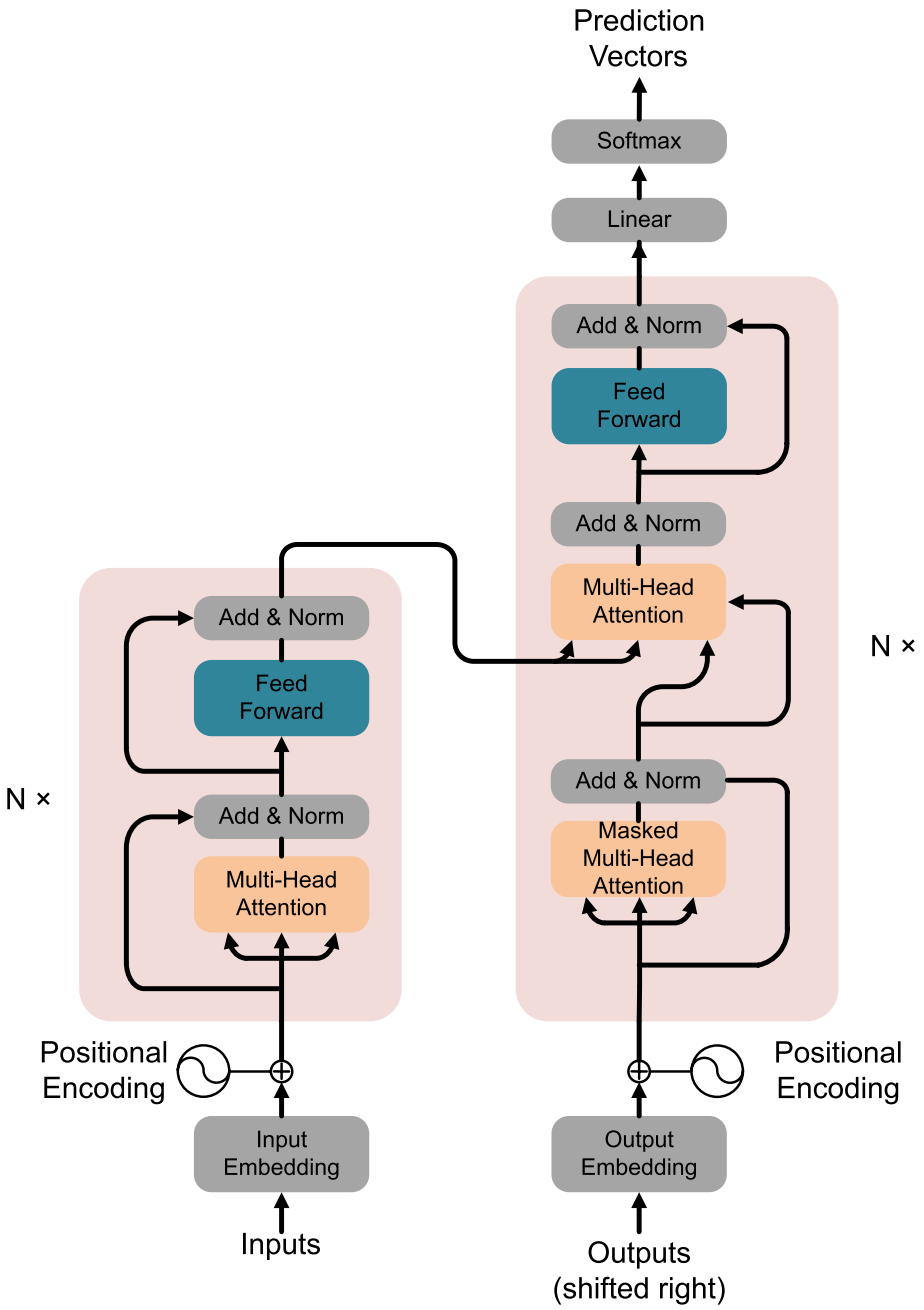


Fig. 7 The transformer model (Vaswani et al. 2017)

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{30}$$

where d_k is the dimension of queries or keys.

Multi-head attention To jointly consider the information from different subspaces of embedding, query, key, and value vectors are mapped into h vectors of identical shapes by using different linear transformations, where h denotes the number of heads. Attention is computed on each of these vectors in parallel, and the results are concatenated and further projected. The multi-head attention can be described as:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{31}$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ and W denotes the linear transformations.

Positional encoding The proposed transformer architecture has no recurrent units, which means that the order information of the sequence is dismissed. The positional encoding is added with input embeddings to provide positional information. The paper chooses cosine functions for positional encoding:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{32}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{33}$$

where *pos* denotes the position of the target token and i denotes the dimension, which means that each dimension of the positional matrix uses a different wavelength for encoding.

Transformer-based pretrain models and transformer variants Recently, many transformer-based pretrain models have been developed. Unlike Embeddings from Language Model (ELMo) proposed by Peters et al. (2018), which is an LSTM-based contextual embedding model, transformer-based pretrain models are more powerful. Two most popular models are GPT-2[5] and BERT (Devlin et al. 2019). GPT-2 and BERT both consist of 12 transformer blocks and BERT is further improved by making the training bi-directional. They are powerful due to their capability of adapting to new tasks after pretraining. This property helped achieve significant improvements in many NLP tasks. There also evolve many Transformer variants (Zaheer et al. 2020; Dai et al. 2019; Guo et al. 2019), which are designed to reduce the model parameters/computational complexity or improve the performance of the original Transformer in diverse scenarios. Lin et al. (2021) and Tay et al. (2020) systematically summarize the state-of-the-art Transformer variants for academics that are interested.

Attention for dialogue systems Attention is a mechanism to catch the importance of different parts in the target sequence. Zhu et al. (2019) applied a two-level attention to generate words. Given the user message and candidate responses selected by a retrieval system, the generator first computes word-level attention weights, then uses sentence-level attention to rescale the weights. This two-level attention helps the generator catch different importance given the encoded context. Liu et al. (2019) used an attention-based recurrent architecture to generate responses. They designed a multi-level encoder-decoder of which the multi-level encoder tries to map raw words, low-level clusters, and high-level clusters into hierarchical embedded representations while the multi-level decoder leveraged the hierarchical representations using attention and then generated responses. At each decoding stage, the model calculated two attention weights for the output of the higher-level

[5] https://openai.com/blog/better-language-models/.

decoder and the hidden state of the current level's encoder. Chen et al. (2019a) computed multi-head self-attention for the outputs of a dialogue act predictor. Unlike the transformer, which concatenates the outputs of different heads, they passed the outputs directly to the next multi-head layer. The stacked multi-head layers then generated the responses with dialogue acts as the input.

Transformers for dialogue systems Transformers are powerful sequence-to-sequence models and meanwhile, their encoders also serve as good dialogue representation models. Henderson et al. (2019b) built a transformer-based response retrieval model for task-oriented dialogue systems. A two-channel transformer encoder was designed for encoding user messages and responses, both of which were initially presented as unigrams and bigrams. A simple cosine distance was then applied to calculate the semantic similarity between the user message and the candidate response. Li et al. (2019c) built multiple incremental transformer encoders to encode multi-turn conversations and their related document knowledge. The encoded utterance and related document of the previous turn were treated as a part of the input of the next turn's transformer encoder. The pretrained model was adaptable to multiple domains with only a small amount of data from the target domain. Bao et al. (2020) used stacked transformers for dialogue generation pretraining. Besides the response generation task, they also pretrained the model together with a latent act prediction task. A latent variable was applied to solve the "one-to-many" problem in response generation. The multi-task training scheme improved the performance of the proposed transformer pretraining model.

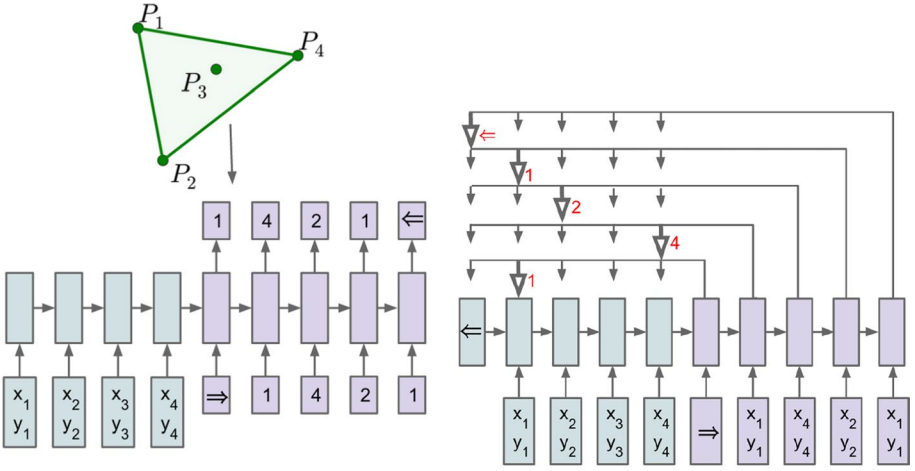
Transformer-based pretrain models for dialogue systems Large transformer-based pretrain models are adaptable to many tasks and are thus popular in recent works. Golovanov et al. (2019) used GPT as a sequence-to-sequence model to directly generate utterances and compared the performances under single- and multi-input settings. Majumder et al. (2020b) first used a probability model to retrieve related news corpus and then combined the news corpus and dialogue context as input of a GPT-2 generator for response generation. They proposed that by using discourse pattern recognition and interrogative type prediction as two subtasks for multi-task learning, the dialogue modeling could be further improved. Wu et al. (2019c) used BERT as an encoder of context and candidate responses in their goal-based response retrieval system while Zhong et al. (2020) built Co-BERT, a BERT-based response selection model, to retrieve empathetic responses given persona-based training corpus. Zhao et al. (2020b) built a knowledge-grounded dialogue system in a synthesized fashion. They used both BERT and GPT-2 to perform knowledge selection and response generation jointly, where BERT was for knowledge selection and GPT-2 generated responses based on dialogue context and the selected knowledge.

2.6 Pointer Net and CopyNet

2.6.1 Pointer Net

In some NLP tasks like dialogue systems and question-answering, the agents sometimes need to directly quote from the user message. Pointer Net (Vinyals et al. 2015) (Fig. 8) solved the problem of directly copying tokens from the input sentence.

Traditional sequence-to-sequence models (Sutskever et al. 2014; Graves et al. 2014) with an encoder–decoder structure map a source sentence to a target sentence. Generally, these models first map the source sentence into hidden state vectors with an encoder and then predict the output sequence based on the hidden states. The sequence prediction is



(a) Sequence-to-sequence **(b)** Pointer Net

Fig. 8 **a** *Sequence-to-sequence*—The RNN (blue) processes the input sequence to produce a code vector, which is then used by the probability chain rule and another RNN to generate the output sequence (purple). The dimensionality of the problem determines the output dimensionality, which remains constant through training and inference. **b** *Pointer Net*—The input sequence is converted to a code (blue) by an encoding RNN, which is fed to the generating network (purple). The generating network generates a vector at each step that modulates a content-based attention process across inputs. The attention mechanism produces a softmax distribution with a dictionary size equal to the input length (Vinyals et al. 2015)

accomplished step-by-step, each step predicting one token using greedy search or beam search. The overall sequence-to-sequence model can be described by the following probability model:

$$P(C^P|P;\theta) = \prod_{i=1}^{m(P)} p(C_i|C_1, \ldots, C_{i-1}, P;\theta) \tag{34}$$

where (P, C_p) constitutes a training pair, $P = \{P_1, \ldots, P_n\}$ denotes the input sequence and $C_p = \{C_1, \ldots, C_{m(p)}\}$ denotes the ground target sequence. θ is a decoder model.

The sequence-to-sequence models have the vanilla backbones and attention-based backbones. Vanilla models predict the target sequence based only on the last hidden state of the encoder and pass it across different decoder time steps. Such a mechanism restricts the information received by the decoder at each decoding stage. Attention-based models consider all hidden states of the encoder at each decoding step and calculate their importance when utilizing them. To compare the mechanism of Pointer Net and Attention, we present the equations explained in Sect. 2.2 here again. The decoder predicts the token conditioned partially on the weighted sum of encoder hidden states d_i:

$$d_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{35}$$

where α_{ij} is the normalized weight score:

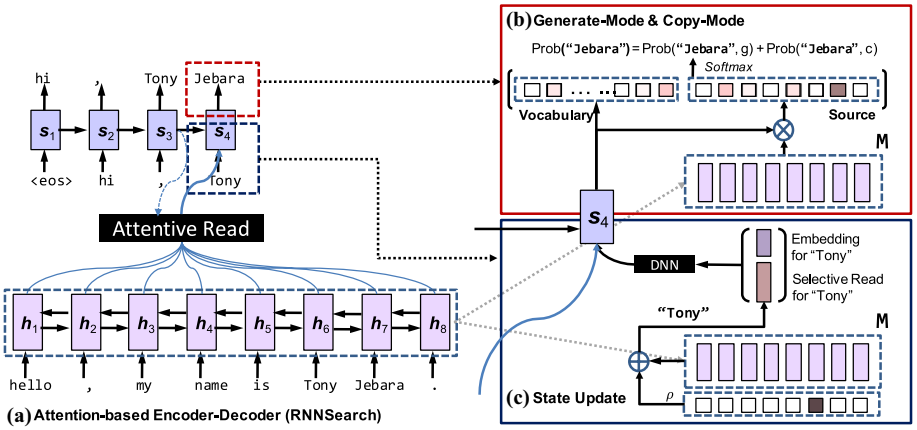


Fig. 9 The overall architecture of CopyNet (Gu et al. 2016)

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{36}$$

e_{ij} is the similarity score between s_{i-1} and *jth* encoder hidden state h_j, where the score is predicted by the similarity model a:

$$e_{ij} = a(s_{i-1}, h_j) \tag{37}$$

At each decoding step, both vanilla and attention-based sequence-to-sequence models predict a distribution over a fixed dictionary $X = \{x_1, \ldots, x_n\}$, where x_i denotes the tokens and n denotes the total count of different tokens in the training corpus. However, when copying words from the input sentence, we do not need such a large dictionary. Instead, n equals to the number of tokens in the input sequence (including repeated ones) and is not fixed since it changes according to the length of the input sequence. Pointer Net made a simple change to the attention-based sequence-to-sequence models: instead of predicting the token distribution based on the weighted sum of encoder hidden states d_i, it directly used the normalized weights α_i as predicted distribution:

$$P(C_i|C_1, \ldots, C_{i-1}, P) = \alpha_i \tag{38}$$

where α_i is a set of probability numbers $\{\alpha_i^1, \ldots, \alpha_i^j\}$ which represents the probability distribution over the tokens of the input sequence. Obviously, the *token prediction* problem is now transformed into *position prediction* problem, where the model only needs to predict a position in the input sequence. This mechanism is like a pointer that points to its target, hence the name "Pointer Net".

2.6.2 CopyNet

In real-world applications, simply copying from the source message is not enough. Instead, in tasks like dialogue systems and QA, agents also require the ability to generate words that are not in the source sentence. CopyNet (Gu et al. 2016) (Fig. 9) was proposed to

incorporate the copy mechanism into traditional sequence-to-sequence models. The model decides at each decoding stage whether to copy from the source or generate a new token not in the source.

The encoder of CopyNet is the same as that of a traditional sequence-to-sequence model, whereas the decoder has some differences compared with a traditional attention-based decoder. When predicting the token at time step t, it combines the probabilistic models of generate-mode and copy-mode:

$$P(y_t|s_t, y_{t-1}, c_t, M) = P_g(y_t|s_t, y_{t-1}, c_t, M) + P_c(y_t|s_t, y_{t-1}, c_t, M) \tag{39}$$

where t is the time step. s_t is the decoder hidden state and y_t is the predicted token. c_t and M represent weighted sum of encoder hidden states and encoder hidden states respectively. g and c are generate-mode and copy-mode respectively.

Besides, though it still uses y_{t-1} and weighted attention vector c_t to update the decoder hidden state, y_{t-1} is uniquely encoded with both its embedding and its location-specific hidden state; also, CopyNet combines attentive read and selective read to capture information from the encoder hidden states, where the selective read is the same method used in Pointer Net. Different from the Neural Turing Machines (Graves et al. 2014; Kurach et al. 2016), the CopyNet has a location-based mechanism that enables the model to be aware of some specific details in training data in a more subtle way.

Copy mechanism is suitable for dialogues involving terminologies or external knowledge sources, and it is popular in knowledge-grounded or task-oriented dialogue systems.

Copy mechanism for knowledge-grounded dialogue systems For knowledge-grounded systems, external documents or dialogues are sources to copy from. Lin et al. (2020a) combined a recurrent knowledge interactive decoder with a knowledge-aware pointer network to achieve both knowledge-grounded generation and knowledge copy. In the proposed model, they first calculated the attention distribution over external knowledge, then used two pointers referring to dialogue context and knowledge source respectively to copy out-of-vocabulary (OOV) words. Wu et al. (2020b) applied a multi-class classifier to flexibly fuse three distributions: generated words, generated knowledge entities, and copied query words. They used Context-Knowledge Fusion and Flexible Mode Fusion to perform the knowledge retrieval, response generation, and copying jointly, making the generated responses precise, coherent, and knowledge-infused. Ji et al. (2020) proposed a Cross Copy Network to copy from internal utterance (dialogue history) and external utterance (similar cases) respectively. They first used pretrained language models for similar case retrieval, then combined the probability distribution of two pointers to make a prediction. They only experimented with court debate and customer service content generation tasks, where similar cases were easy to obtain.

Copy mechanism for task-oriented dialogue systems Many dialogue state tracking tasks generate slots and slot values using a copy component (Wu et al. 2019a; Ouyang et al. 2020; Gangadharaiah and Narayanaswamy 2020; Chen et al. 2020a; Zhang et al. 2020b; Li et al. 2020e). Among them Wu et al. (2019a), Ouyang et al. (2020) and Chen et al. (2020a) solved the problem of multi-domain dialogue state tracking. Wu et al. (2019a) proposed TRAnsferable Dialogue statE generator (TRADE), a copy-based dialogue state generator. The generator decoded the slot value multiple times for each possible (domain, slot) pair, then a slot gate was applied to decide which pair belonged to the dialogue. The output distribution was a copy of the slot values belonging to the selected (domain, slot) pairs from vocabulary and dialogue history. Chen et al. (2020a) used a different copy strategy from TRADE. Instead of using the whole dialogue history as the copy source, they copied state

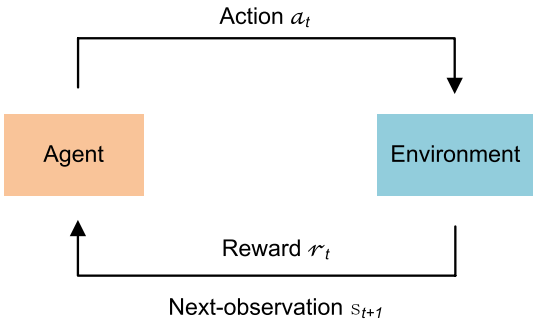


Fig. 10 The reinforcement learning framework

values from user utterances and system messages respectively, which took the slot-level context as input. Ouyang et al. (2020) proposed slot connection mechanism to efficiently utilize existing states from other domains. Attention weights were calculated to measure the connection between the target slot and related slot-value tuples in other domains. Three distributions over token generation, dialogue context copying, and past state copying were finally gated and fused to predict the next token. Gangadharaiah and Narayanaswamy (2020) combined a pointer network with a template-based tree decoder to fill the templates recursively and hierarchically. Copy mechanisms also alleviated the problem of expensive data annotation in end-to-end task-oriented dialogue systems. Copy-augmented dialogue generation models were proven to perform significantly better than strong baselines with limited domain-specific or multi-domain data (Zhang et al. 2020b; Li et al. 2020e; Gao et al. 2020a).

Copy mechanism for dialogue-related tasks Pointer networks and CopyNet are also used to solve other dialogue-related tasks. Yu and Joty (2020) applied a pointer net for online conversation disentanglement. The pointer module pointed to the ancestor message to which the current message replies and a classifier predicted whether two messages belonged to the same thread. In dialogue parsing tasks, the pointer net is used as the backbone parsing model to construct discourse trees (Aghajanyan et al. 2020; Lin et al. 2019). Tay et al. (2019) used a pointer-generator framework to perform machine reading comprehension over a long span, where the copy mechanism reduced the demand for including target answers in context.

2.7 Deep reinforcement learning models and generative adversarial networks

In recent years, two exciting approaches have exhibited the potential of artificial intelligence. The first one is deep reinforcement learning, which outperforms humans in many complex problems such as large-scale games, conversations, and car-driving. Another technique is GAN, showing amazing capability in generation tasks. The data samples generated by GAN models like articles, paintings, and even videos, are sometimes indistinguishable from human creations.

AlphaGo (Silver et al. 2016) stimulated the research interests again in reinforcement learning in recent years (Graves et al. 2016; Mnih et al. 2016; Wang et al. 2016; Tamar et al. 2016; Jaderberg et al. 2017; Mirowski et al. 2017). Reinforcement learning is a branch of machine learning aiming to train agents to perform appropriate actions while interacting with a certain environment. It is one of the three fundamental machine learning branches,

with supervised learning and unsupervised learning being the other two. It can also be seen as an intermediate between supervised learning and unsupervised learning because it only needs weak signals for training (Wang et al. 2016).

Figure 10 illustrates the reinforcement learning framework, consisting of an agent and an environment. The framework is a Markov Decision Process (MDP) (Puterman 2014), which can be described by a five-tuple $\mathrm{M} = \langle S, A, P, R, \gamma \rangle$. S denotes an infinite set of environment states; A denotes a set of actions that an agent chooses from conditioned on a given environment state s; P is the transition probability matrix in MDP, denoting the probability of an environment state transfer after agent takes an action; R is an average reward the agent receives from the environment after taking an action under state s; γ is a discount factor. The flow of this framework is a loop of the following two steps: the agent first makes an observation of the current environment state s_t and chooses an action based on its policy; then according to the transition probability matrix P, the environment's state transfers to s_{t+1}, and simultaneously provides a reward r_t.

Reinforcement learning is applicable to solve many challenges in dialogue systems because of the agent-environment nature of a dialogue system. A two-party dialogue system consists of an agent, which is an intelligent chatbot, and an environment, which is usually a user or a user simulator. Here we mainly discuss deep reinforcement learning.

Deep reinforcement learning means applying deep neural networks to model the value function or policy of the reinforcement learning framework. “Deep model" is in contrast to the “shallow model". The shallow model normally refers to traditional machine learning models like Decision Trees or KNN. Feature engineering, which is usually based on shallow models, is time and labor-consuming, and also over-specified and incomplete. Different from that, deep neural models are easy to design and have a strong fitting capability, which contributes to many breakthroughs in recent research. Deep representation learning gets rid of human labor and exploits hierarchical features in data automatically, which strengthens the semantic expressiveness and domain correlations significantly.

We discuss two typical reinforcement models: Deep Q-Networks (Mnih et al. 2015) and REINFORCE (Williams 1992; Sutton et al. 1999). They belong to *Q-learning* and *policy gradient* respectively, which are two families of reinforcement learning.

2.7.1 Deep Q-Networks

A Deep Q-Network is a value-based RL model. It determines the best policy according to the Q-function:

$$\pi^*(s) = arg\max_a Q^*(s, a) \tag{40}$$

where $Q^*(s, a)$ is an optimal Q-function and $\pi^*(s)$ is the corresponding optimal policy. In Deep Q-Networks, the Q function is modeled using a deep neural network, such as CNNs, RNNs, etc.

As in Gao et al. (2018), the parameters of the Q model are updated using the rule:

$$\theta \leftarrow \theta + \alpha \underbrace{\left(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta) - Q(s_t, a_t; \theta) \right)}_{\text{temporal difference}} \nabla_\theta Q(s_t, a_t; \theta) \tag{41}$$

where the (s_t, a_t, r_t, s_{t+1}) is an observed trajectory. α denotes step-size and the parameter update is calculated using temporal difference (Sutton 1988). However, this update mechanism suffers from unstableness and demands a large number of training samples. There are two typical tricks for a more efficient and stable parameter update.

The first method is experience replay (Lin 1992; Mnih et al. 2015). Instead of using one training sample at a time to update the parameters, it uses a buffer to store training samples, and iteratively retrieves training samples from the buffer pool to perform parameter updates. It avoids encountering training samples that change too fast in distribution during training time, which increases the learning stability; further, it uses each training sample multiple times, which improves the efficiency.

The second is a two-network implementation (Mnih et al. 2015). This method uses two networks in Q-function optimization, one being the Q-network, and another being a target network. The target network is used to calculate the temporal difference, and its parameters θ_{target} are frozen while training, aligning with θ periodically. The parameters are then updated with the following rule:

$$\theta \leftarrow \theta + \alpha \underbrace{\left(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_{target}) - Q(s_t, a_t; \theta) \right)}_{\text{temporal difference with a target network}} \nabla_\theta Q(s_t, a_t; \theta) \tag{42}$$

Since θ_{target} does not change in a period of time, the target network calculates the temporal difference in a stable manner, which facilitates the convergence of training.

2.7.2 REINFORCE

REINFORCE is a policy-based RL algorithm that has no value network. It optimizes the policy directly. The policy is parameterized by a policy network, whose output is a distribution over continuous or discrete actions. A long-term reward is computed for evaluation of the policy network by collecting trajectory samples of length H:

$$J(\theta) = E\left[\sum_{t=1}^{H} \gamma^{t-1} r_t | a_t \sim \pi(s_t; \theta) \right] \tag{43}$$

$J(\theta)$ denotes a long-term reward and the goal is to optimize the policy network in order to maximize $J(\theta)$. Here stochastic gradient ascent[6] is used as an optimizer:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta) \tag{44}$$

where $\nabla_\theta J(\theta)$ is computed by:

$$\nabla_\theta J(\theta) = \sum_{t=1}^{H-1} \gamma^{t-1} \left(\nabla_\theta log \pi(a_t | s_t; \theta) \sum_{h=t}^{H} \gamma^{h-t} r_h \right) \tag{45}$$

Both models have their advantages: Deep Q-Networks are more sample efficient while REINFORCE is more stable (Li 2017). REINFORCE is more popular in recent works. Modern research involves larger action spaces, which means that value-based RL models

[6] Stochastic gradient ascent simply uses the negated objective function of stochastic gradient descent.

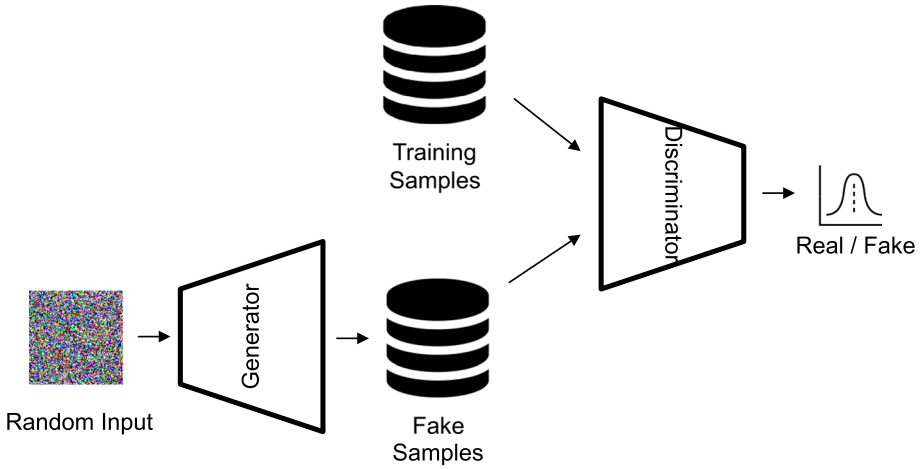


Fig. 11 The GAN framework

like Deep Q-Networks are not suitable for problem-solving. Value-based methods "select an action to maximize the value", which means that their action sets should be discrete and moderate in scale; while policy gradient methods such as REINFORCE are different, they predict the action via policy networks directly, which sets no restriction on the action space. As a result, policy gradient methods are more suitable for tasks involving a larger action space.

Considering the respective benefits brought by the Q-learning and policy gradient, some work has been done combining the value- and policy-based methods. Actor-critic algorithm (Konda and Tsitsiklis 2000; Sutton et al. 1999) was proposed to alleviate the severe variance problem when calculating the gradient in policy gradient methods. It estimates a value function for term $\sum_{h=t}^{H} \gamma^{h-t} r_h$ in Eq. (45) and incorporates it in policy optimization. Equation (45) is then transformed into the formula below:

$$\nabla_\theta J(\theta) = \sum_{t=1}^{H-1} \gamma^{t-1} \left(\nabla_\theta log \pi(a_t|s_t;\theta) \hat{Q}(s_t, a_t, h) \right) \tag{46}$$

Where $\hat{Q}(s_t, a_t, h)$ stands for the value function estimated.

2.7.3 GANs

It is easy to link the actor-critic model with another framework - GANs (Goodfellow et al. 2014; Zhang et al. 2018b; Feng et al. 2020a) because of their similar inner structure and logic (Pfau and Vinyals 2016). Actually, there are quite a few recent works in dialogue systems that train GANs with reinforcement learning framework (Zhu et al. 2019; Wu et al. 2019b; He et al. 2020a; Zhu et al. 2020; Qin et al. 2020).

Figure 11 represents the GAN consisting of a generator and a discriminator where the training process can be viewed as a competition between them: the generator tries to generate data distributions to fool the discriminator while the discriminator attempts to distinguish between real data (real) and generated data (fake). During training, the generator

takes noise as input and generates data distribution while the discriminator takes real and fake data as input and the binary annotation as the label. The whole GAN model is trained end-to-end as a connection of generator and discriminator to minimize the following cross-entropy losses:

$$L_1(D, G) = -E_{\omega \sim P_{data}}[logD(\omega)] - E_{z \sim N(0,I)}[log(1 - D(G(z)))] \tag{47}$$

$$L_2(D, G) = -E_{z \sim N(0,I)}[logD(G(z))] \tag{48}$$

where L_1 and L_2 denote a bilevel loss, where D and G are discriminator and generator respectively. $z \sim N(0, I)$ is the noise input of the generator and w is the input of the discriminator.

Relationship between RL and GAN GAN can be viewed as a special actor-critic (Pfau and Vinyals 2016). In the learning architecture of GAN, the generator acts as the actor and the discriminator acts as the critic or environment which gives the real/fake feedback as a reward. However, the actions taken by the actor cannot change the states of the environment, which means that the learning architecture of GAN is a stateless Markov decision process. Also, the actor has no access to the state of the environment and generates data distribution simply conditioned on Gaussian noise, which means that the generator in the GAN framework is a blind actor/agent. In a nutshell, GAN is a special actor-critic where the actor is blind and the whole process is a stateless MDP.

The interactive nature of dialogue systems motivates the wide application of reinforcement learning and GAN models in its research.

RL for task-oriented dialogue systems One common application of reinforcement learning in dialogue systems is the reinforced dialogue management in task-oriented systems. Dialogue state tracking and policy learning are two typical modules of a dialogue manager. Huang et al. (2020c) and Li et al. (2020e) trained the dialogue state tracker with reinforcement learning. Both of them combined a reward manager into their tracker to enhance tracking accuracy. For the policy learning module, reinforcement learning seems to be the best choice since almost all recent related works learned policy with reinforcement learning (Zhang et al. 2019d; Wang et al. 2020d; Zhu et al. 2020; Wang et al. 2020a; Takanobu et al. 2020; Huang et al. 2020b; Xu et al. 2020a). The increasing preference for reinforcement learning in policy learning tasks attributes to the characteristic of them: in policy learning tasks, the model predicts a dialogue action (action) based on the states from the DST module (state), which perfectly accords with the function of the agent in the reinforcement learning framework.

RL for open-domain dialogue systems Due to the huge action space needed to generate language directly, many open-domain dialogue systems trained with reinforcement learning framework do not generate responses but instead select responses. Retrieval-based systems have a limited action set and are suitable to be trained in a reinforcement learning scheme. Some works achieved promising performance in retrieval-based dialogue tasks (Bouchacourt and Baroni 2019; Li et al. 2017a; Zhao and Eskenazi 2016). However, retrieval systems fail to generalize in all user messages and may give unrelated responses (Qiu et al. 2017), which makes generation-based dialogue systems preferable. Still considering the action space problem, some works build their systems combining retrieval and generative methods (Zhu et al. 2019; Serban et al. 2017a). Zhu et al. (2019) chose to first retrieve a set of n-best response candidates and then generated responses based on the retrieved results and user message. Comparatively, Serban et al. (2017a) first generated and retrieved candidate responses with different dialogue models and then trained a scoring model with online

reinforcement learning to select responses from both generated and retrieved responses. Since training a generative dialogue agent using reinforcement learning from scratch is particularly difficult, first pretraining the agent with supervised learning to warm-start is a good choice. Wu et al. (2019b), He et al. (2020a), Williams and Zweig (2016) and Yao et al. (2016) applied this pretrain-and-finetune strategy on dialogue learning and achieved outstanding performance, which proved that the reinforcement learning can improve the response quality of data-driven chatbots. Similarly, pretrain-and-finetune was also applicable to domain transfer problems. Some works pretrained the model in a source domain and expanded the domain area with reinforcement training (Mo et al. 2018; Li et al. 2016c).

RL for knowledge grounded dialogue systems Some systems use reinforcement learning to select from outside information like persona, document, knowledge graph, etc., and generate responses accordingly. Majumder et al. (2020a) and Jaques et al. (2020) performed persona selection and persona-based response generation simultaneously and trained their agents with a reinforcement framework. Bao et al. (2019) and Zhao et al. (2020b) built document-grounded systems. Similarly, they used reinforcement learning to accomplish document selection and knowledge-grounded response generation. There were also some works combining knowledge graphs into the dialogue systems and treating them as outside knowledge source (Moon et al. 2019; Xu et al. 2020a). In a reinforced training framework, the agent chooses an edge based on the current node and state for each step and then combines the knowledge into the response generation process.

RL for dialogue related tasks Dialogue-related tasks like dialogue relation extraction (Li et al. 2019b), question answering (Hua et al. 2020) and machine reading comprehension (Guo et al. 2020) benefit from reinforcement learning as well because of their interactive nature and the scarcity of annotated data.

GAN for dialogue systems The application of GAN in dialogue systems is divided into two streams. The first sees the GAN framework applied to enhance response generation (Li et al. 2017c; Zhu et al. 2019; Wu et al. 2019b; He et al. 2020a; Zhu et al. 2020; Qin et al. 2020). The discriminator distinguishes generated responses from human responses, which incentivizes the agent, which is also the generator in GAN, to generate higher-quality responses. Another stream uses GAN as an evaluation tool for dialogue systems (Kannan and Vinyals 2017; Bruni and Fernández 2017). After training the generator and discriminator as a whole framework, the discriminator is used separately as a scorer to evaluate the performance of a dialogue agent and was shown to achieve a higher correlation with human evaluation compared with traditional reference-based metrics like BLEU, METEOR, ROUGE-L, etc. We discuss the evaluation of dialogue systems as a challenge in Sect. 5.

2.8 Knowledge graph augmented neural networks

Supervised training with annotated data tries to learn the knowledge distribution of a dataset. However, a dataset is comparatively sparse and thus learning a reliable knowledge distribution needs a huge amount of annotated data (K M et al. 2018).

Knowledge Graph (KG) is attracting more and more research interests in recent years. KG is a structured knowledge source consisting of entities and their relationships (Ji et al. 2022). In other words, KG is the knowledge facts presented in graph format.

Figure 12 shows an example of a KG consisting of entities and their relationships. A KG is stored in triples under the Resource Description Framework (RDF). For example, Albert Einstein, University of Zurich, and their relationship can be expressed as (*AlbertEinstein*, *GraduateFrom*, *UniversityofZurich*).

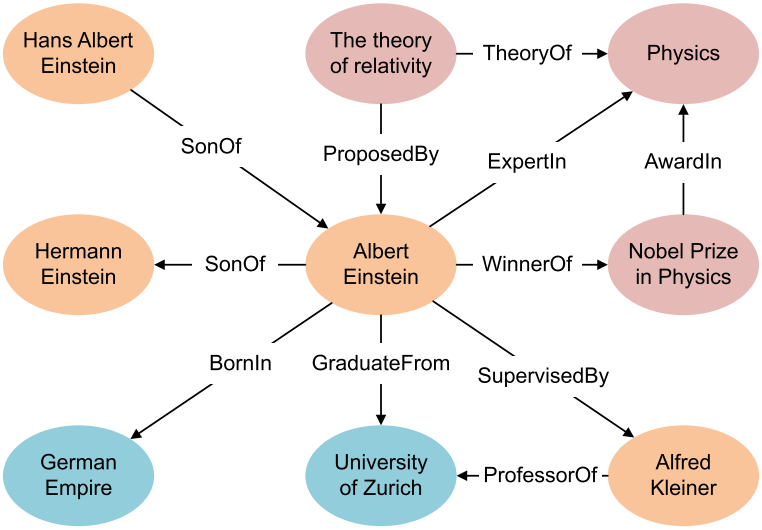


Fig. 12 Entities and relations in knowledge graph (Ji et al. 2022)

Knowledge graph augmented neural networks first represent the entities and their relations in a lower dimension space, then use a neural model to retrieve relevant facts (Ji et al. 2022). Knowledge graph representation learning can be generally divided into two categories: structure-based representations and semantically-enriched representations. Structure-based representations use multi-dimensional vectors to represent entities and relations. Models such as TransE (Bordes et al. 2013), TransR (Lin et al. 2015), TransH (Wang et al. 2014), TransD (Ji et al. 2015), TransG (Xiao et al. 2015), TransM (Fan et al. 2014), HolE (Nickel et al. 2016) and ProjE (Shi and Weninger 2017) belong to this category. The semantically-enriched representation models like NTN (Socher et al. 2013), SSP (Xiao et al. 2017) and DKRL (Xie et al. 2016) combine semantic information into the representation of entities and relations. The neural retrieval models also have two main directions: distance-based matching model and semantic matching model. Distance-based matching models (Bordes et al. 2013) consider the distance between projected entities while semantic matching models (Bordes et al. 2014) calculate the semantic similarity of entities and relations to retrieve facts.

Knowledge graph augmented dialogue systems Knowledge-grounded dialogue systems benefit greatly from the structured knowledge format of KG, where facts are widely intercorrelated. Reasoning over a KG is an ideal approach for combining commonsense knowledge into response generation, resulting in accurate and informative responses (Young et al. 2018). Jung et al. (2020) proposed AttnIO, a bi-directional graph exploration model for knowledge retrieval in knowledge-grounded dialogue systems. Attention weights were calculated at each traversing step, and thus the model could choose a broader range of knowledge paths instead of choosing only one node at a time. In such a scheme, the model could predict adequate paths even when only having the destination node as the label. Zhang et al. (2020a) built ConceptFlow, a dialogue agent that guided to more meaningful future conversations. It traversed in a commonsense knowledge graph to explore concept-level conversation flows. Finally, it used a gate to decide to generate among vocabulary words, central concept words, and outer concept words. Majumder et al. (2020a) proposed to

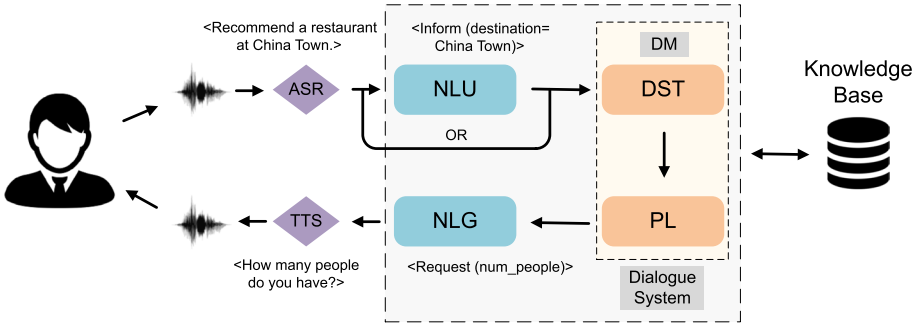


Fig. 13 Structure of a task-oriented dialogue system in the task-completion pipeline

generate persona-based responses by first using COMET (Bosselut et al. 2019) to expand a persona sentence in context along 9 relation types and then applied a pretrained model to generate responses based on dialogue history and the persona variable. Yang et al. (2020) used knowledge graph as an external knowledge source in task-oriented dialogue systems to incorporate domain-specified knowledge in the response. First, the dialogue history was parsed as a dependency tree and encoded into a fixed-length vector. Then they applied multi-hop reasoning over the graph using the attention mechanism. The decoder finally predicted tokens either by copying from graph entities or generating vocabulary words. Moon et al. (2019) proposed DialKG Walker for the conversational reasoning task. They computed a zero-shot relevance score between predicted KG embedding and ground KG embedding to facilitate cross-domain predictions. Furthermore, they applied an attention-based graph walker to generate graph paths based on the relevance scores. Huang et al. (2020a) evaluated the dialogue systems by combining the utterance-level contextualized representation and topic-level graph representation. They first constructed the dialogue graph based on encoded (context, response) pairs and then reasoned over the graph to get a topic-level graph representation. The final score was calculated by passing the concatenated vector of contextualized representation and graph representation to a feed-forward network.

3 Task-oriented dialogue systems

This section introduces task-oriented dialogue systems including modular and end-to-end systems. Task-oriented systems solve specific problems in a certain domain such as movie ticket booking, restaurant table reserving, etc. We focus on deep learning-based systems due to the outstanding performance. For readers who want to learn more about traditional rule-based and statistical models, there are several surveys to refer to (Theune 2003; Lemon and Pietquin 2007; Mallios and Bourbakis 2016; Chen et al. 2017a; Santhanam and Shaikh 2019).

This section is organized as follows. We first discuss modular and end-to-end systems respectively by introducing the principles and reviewing recent works. After that, we comprehensively discuss related challenges and hot topics for task-oriented dialogue systems in recent research to provide some important research directions.

Table 2 The output example of an NLU module

Sentence	Recommend	a	movie	at	Golden	Village	tonight
Slots	O	O	O	O	B-desti	I-desti	B-time
Intent				find_movie			
Domain				movie			

A task-oriented dialogue system requires stricter response constraints because it aims to accurately handle the user message. Therefore, modular methods were proposed to generate responses in a more controllable way. The architecture of a modular-based system is depicted in Fig. 13. It consists of four modules:

Natural language understanding (NLU) This module converts the raw user message into semantic slots, together with classifications of domain and user intention. However, some recent modular systems omit this module and use the raw user message as the input of the next module, as shown in Fig. 13. Such a design aims to reduce the propagation of errors between modules and alleviate the impact of the original error (Kim et al. 2018).

Dialogue state tracking (DST) This module iteratively calibrates the dialogue states based on the current input and dialogue history. The dialogue state includes related user actions and slot-value pairs.

Dialogue policy learning Based on the calibrated dialogue states from the DST module, this module decides the next action of a dialogue agent.

Natural language generation (NLG) This module converts the selected dialogue actions into surface-level natural language, which is usually the ultimate form of response.

Among them, Dialogue State Tracking and Dialogue Policy Learning constitute the Dialogue Manager (DM), the central controller of a task-oriented dialogue system. Usually, a task-oriented system also interacts with an external Knowledge Base (KB) to retrieve essential knowledge about the target task. For example, in a movie ticket booking task, after understanding the requirement of the user message, the agent interacts with the movie knowledge base to search for movies with specific constraints such as movie name, time, cinema, etc.

3.1 Natural language understanding

It has been proven that the NLU module impacts the whole system significantly in the term of response quality (Li et al. 2017d). The NLU module converts the natural language message produced by the user into semantic slots and performs classification. Table 2 shows an example of the output format of the NLU module. The NLU module manages three tasks: domain classification, intent detection, and slot filling. Domain classification and intent detection are classification problems, which use classifiers to predict a mapping from the input language sequence to a predefined label set. In the given example, the predicted domain is "*movie*" and the intent is "*find_movie*". Slot filling is a tagging problem, which can be viewed as a sequence-to-sequence task. It maps a raw user message into a sequence of slot names. In the example, the NLU module reads

the user message "*Recommend a movie at Golden Village tonight.*" and outputs the corresponding tag sequence. It recognizes "*Golden Village*" as the place to go, which is tagged as "*B_desti*" and "*I_desti*" for the two words respectively. Similarly, the token "*tonight*" is converted into "*B_time*". 'B' represents the beginning of a chunk, and 'I' indicates that this tag is inside a target chunk. For those unrelated tokens, an 'O' is used indicating that this token is outside of any chunk of interest. This tagging method is called Inside-Outside-Beginning (IOB) tagging (Ramshaw and Marcus 1995), which is a common method in Named-Entity Recognition (NER) tasks.

Techniques for domain classification and intent detection Domain classification and intent detection belong to the same category of tasks. Deep learning methods are proposed to solve the classification problems of dialogue domain and intent. Deng et al. (2012) and Tur et al. (2012) were the first who successfully improved the recognition accuracy of dialogue intent. They built deep convex networks to combine the predictions of a prior network and the current utterances as an integrated input of a current network. A deep learning framework was also used to classify the dialogue domain and intent in a semi-supervised fashion (Yann et al. 2014). To solve the difficulty of training a deep neural network for domain and intent prediction, Restricted Boltzmann Machine (RBM) and Deep Belief Networks (DBNs) were applied to initialize the parameters of deep neural networks (Sarikaya et al. 2014). To make use of the strengths of RNNs in sequence processing, some works used RNNs as utterance encoders and made predictions for intent and domain categories (Ravuri and Stolcke 2015, 2016). Hashemi et al. (2016) used a CNN to extract hierarchical text features for intent detection and illustrated the sequence classification capabilities of CNNs. Lee and Dernoncourt (2016) proposed a model for intent classification of short utterances. Short utterances are hard for intent detection because of the lack of information in a single dialogue turn. This paper used RNN and CNN architectures to incorporate the dialogue history, thus obtaining the context information as an additional input besides the current turn's message. The model achieved promising performances on three intent classification datasets. More recently, Wu et al. (2020a) pretrained Task-Oriented Dialogue BERT (TOD-BERT) and significantly improved the accuracy in the intent detection sub-task. The proposed model also exhibited a strong capability of few-shot learning and could effectively alleviate the data insufficiency issue in a specific domain. Casanueva et al. (2020) introduced duel sentence encoders for efficient intent detection. Their methods are effective in low-resource situations. Abro et al. (2022) proposed an NLU framework for argumentative dialogue systems in the information-seeking and opinion-building domain. They use a BERT+BiLSTM to inject commonsense knowledge into the framework to better understand the user intent. Obuchowski and Lew (2020) combined transformer with capsule networks, and find their model achieves better performance than original capsule-NLU network implementations.

Techniques for slot filling The slot filling problem is also called semantic tagging, a sequence classification problem. It is more challenging for that the model needs to predict multiple objects at a time. Deep Belief Nets (DBNs) exhibit promising capabilities in the learning of deep architectures and have been applied in many tasks including semantic tagging. Sarikaya et al. (2011) used a DBN-initialized neural network to complete slot filling in the call-routing task. Deoras and Sarikaya (2013) built a DBN-based sequence tagger. In addition to the NER input features used in traditional taggers, they also combined part of speech (POS) and syntactic features as a part of the input. The recurrent architectures benefited the sequence tagging task in that they could keep track of the information along past timesteps to make the most of the sequential information. Yao et al. (2013) first argued that instead of simply predicting words, RNN Language Models (RNN-LMs) could be applied in sequence

tagging. On the output side of RNN-LMs, tag labels were predicted instead of normal vocabularies. Mesnil et al. (2013) and Mesnil et al. (2014) further investigated the impact of different recurrent architectures in the slot filling task and found that all RNNs outperformed the Conditional Random Field (CRF) baseline. As a powerful recurrent model, LSTM showed promising tagging accuracy on the ATIS dataset owing to the memory control of its gate mechanism (Yao et al. 2014). Gangadharaiah and Narayanaswamy (2020) argued that the shallow output representations of traditional semantic tagging lacked the ability to represent the structured dialogue information. To improve, they treated the slot filling task as a template-based tree decoding task by iteratively generating and filling in the templates. Different from traditional sequence tagging methods, Coope et al. (2020) tackled the slot filling task by treating it as a turn-based span extraction task. They applied the conversational pretrained model ConveRT and utilized the rich semantic information embedded in the pretrained vectors to solve the problem of in-domain data insufficiency. The inputs of ConveRT are the requested slots and the utterance, while the output is a span of interest as the slot value.

Unifying domain classification, intent detection, and slot filling Some works choose to combine domain classification, intent detection, and slot filling into a multitask learning framework to jointly optimize the shared latent space. Hakkani-Tür et al. (2016) applied a bi-directional RNN-LSTM architecture to jointly perform three tasks. Liu and Lane (2016) augmented the traditional RNN encoder-decoder model with an attention mechanism to manage intent detection and slot filling. The slot filling applied explicit alignment. Chen et al. (2016) proposed an end-to-end memory network and used a memory module to store user intent and slot values in history utterances. Attention was further applied to iteratively select relevant intent and slot values at the decoding stage. Multi-task learning of three NLU subtasks contributed to the domain scaling and facilitated the zero-shot or few-shot training when transferring to a new domain (Bapna et al. 2017; Lee and Jha 2019). Zhang et al. (2019a) captured the hierarchical structure of dialogue semantics in NLU multi-task learning by applying a capsule-based neural network. With a dynamic routing-by-agreement strategy, the proposed architecture raised the accuracy of both intent detection and slot filling on the SNIPS-NLU and ATIS datasets. Abro et al. (2020) proposed a multi-task learning model based on neural networks and regular expressions (REs) to tackle the low-resource intent determination and slot filling. Qin et al. (2019) proposed a joint model with StackPropagation which can directly use the intent information as input for slot filling. They further perform token-level intent detection to alleviate the error. Similarly, Wang et al. (2018) proposed a Bi-model based RNN semantic frame parsing network structure to utilize the cross-impact to each other when performing joint intent detection and slot-filling. Zhang et al. (2019b) proposed a capsule-based neural network model which performs slot filling and intent detection via a dynamic routing-by-agreement schema.

Novel perspectives More recently, some novel ideas appear in NLU research, which provides new possibilities for further improvements. Traditional NLU modules rely on the text converted from the audio message of the user using the Automatic Speech Recognition (ASR) module. However, Singla et al. (2020) jumped over the ASR module and directly used audio signals as the input of NLU. They found that by reducing the module numbers of a pipeline system, the predictions were more robust since fewer errors were broadcasted. Su et al. (2019b) argued that Natural Language Understanding (NLU) and Natural Language Generation (NLG) were reversed processes. Thus, their dual relationship could be exploited by training with a dual-supervised learning framework. The experiments exhibited improvement in both tasks.

3.2 Dialogue state tracking

Dialogue State Tracking (DST) is the first module of a dialogue manager. It tracks the user's goal and related details every turn based on the whole dialogue history to provide the information based on which the Policy Learning module (next module) decides the agent action to make.

Differences between NLU and DST The NLU and DST modules are closely related. Both NLU and DST perform slot filling for the dialogue. However, they actually play different roles. The NLU module tries to make classifications for the current user message such as the intent and domain category as well as the slot each message token belongs to. For example, given a user message "*Recommend a movie at Golden Village tonight.*", the NLU module will convert the raw message into " *inform*(*domain* = *movie*; *destination* = *GoldenVillage*; *date* = *today*; *time* = *evening*) ", where the slots are usually filled by tagging each word of the user message as described in Sect. 3.1. However, the DST module does not classify or tag the user message. Instead, it tries to find a slot value for each slot name in a pre-existing slot list based on the whole dialogue history. For example, there is a pre-existing slot list " *intent* : _; *domain* : _; *name* : _; *pricerange* : _; *genre* : _; *destination* : _; *date* : _ ", where the underscore behind the colon is a placeholder denoting that this place can be filled with a value. Every turn, the DST module will look up the whole dialogue history up to the current turn and decide which content can be filled in a specific slot in the slot list. If the user message "*Recommend a movie at Golden Village tonight.*" is the only message in a dialogue, then the slot list can be filled as " *intent* : *inform*; *domain* : *movie*; *name* : *None*; *pricerange* : *None*; *genre* : *None*; *destination* : *GoldenVillage*; *date* : *today*", where the slots unspecified by the user up to current turn can be filled with "*None*". To conclude, the NLU module tries to tag the user message while the DST module tries to find values from the user message to fill in a pre-existing form. Some dialogue systems took the output of the NLU module as the input of the DST module (Williams et al. 2013; Henderson et al. 2014a, b), while others directly used raw user messages to track the state (Kim et al. 2020b; Wang et al. 2020e; Hu et al. 2020).

Dialogue State Tracking Challenges (DSTCs), a series of popular challenges in DST, provides benchmark datasets, standard evaluation frameworks, and test-beds for research (Williams et al. 2013; Henderson et al. 2014a, b; Kim et al. 2016, 2017). The DSTCs cover many domains such as restaurants, tourism, etc.

A dialogue state contains all essential information to be conveyed in the response (Henderson 2015). As defined in DSTC2 (Henderson et al. 2014a), the dialogue state of a given dialogue turn consists of informable slots *Sinf* and requestable slots *Sreq*. Informable slots are attributes specified by users to constrain the search of the database while requestable slots are attributes whose values are queried by the user. For example, the serial number of a movie ticket is usually a requestable slot because users seldom assign a specific serial number when booking a ticket. Specifically, the dialogue state has three components:

- *Goal constraint corresponding with informable slots* The constraints can be specific values mentioned by the user in the dialogue or a special value. Special values include *Dontcare* indicating the user's indifference about the slot and *None* indicating that the user has not specified the value in the conversation yet.

Actual input and output	SLU hypotheses and scores	Labels	Example tracker output	Correct?
S: Which part of town? *request(area)*	0.2 inform(food=north_african)	area=north	0.2 food=north_african	✗
	0.1 inform(area=north)		0.1 area=north	✓
			0.7 ()	✗
U: The north uh area *inform(area=north)*		method=byconstraints	0.9 byconstraints 0.1 none	✓
		requested=()	0.0 phone 0.0 address	✓ ✓
S: Which part of town? *request(area)*	0.8 inform(area=north), inform(pricerange=cheap)	area=north pricerange=cheap	0.7 area=north pricerange=cheap	✓
	0.1 inform(area=north)		0.1 area=north food=north_african	✗
U: A cheap place in the north *inform(area=north, pricerange=cheap)*			0.2 ()	✗
		method=byconstraints	0.9 byconstraints 0.1 none	✓
		requested=()	0.0 phone 0.0 address	✓ ✓
S: Clown café is a cheap restaurant in the north part of town.	0.7 reqalts(area=south)	area=south pricerange=cheap	0.8 area=south pricerange=cheap	✓
	0.2 reqmore()		0.1 area=north pricerange=cheap	✗
U: Do you have any others like that, maybe in the south part of town? *reqalts(area=south)*			0.1 ()	✗
		method=byalternatives	0.6 byalternatives 0.2 byconstraints	✓
		requested=()	0.0 phone 0.0 address	✓ ✓
S: Galleria is a cheap restaurant in the south.	0.6 request(phone)	area=south pricerange=cheap	0.9 area=south pricerange=cheap	✓
	0.2 request(phone), request(address)		0.1 area=north pricerange=cheap	✗
U: What is their phone number and address? *request(phone), request(address)*	0.1 request(address)		0.0 ()	✗
		method=byalternatives	0.5 byconstraints 0.4 byalternatives	✗
		requested= (phone, address)	0.8 phone 0.3 address	✓ ✗

Fig. 14 An example of DST procedure (Henderson et al. 2014a)

- *Requested slots* It can be a list of slot names queried by the user seeking answers from the agent.
- *Search method of current turn* It consists of values indicating the interaction categories. *By constraints* denotes that the user tries to specify constraint information in his requirement; *by alternatives* denotes that the user requires an alternative entity; *finished* indicates that the user intends to end the conversation.

However, considering the numerous challenges such as tracking efficiency, tracking accuracy, domain adaptability, and end-to-end training, many alternative representations have been proposed recently, which will be discussed later.

Figure 14 is an example of the DST process for 4 dialogue turns in a restaurant table booking task. The first column includes the raw dialogue utterances, with *S* denoting the system message and *U* denoting the user message. The second column includes the N-best output lists of the NLU module and their corresponding confidence scores. The third

column includes the labels of a turn, indicating the ground truth slot-value pairs. The fourth column includes the example DST outputs and their corresponding confidence scores. The fifth column indicates the correctness of the tracker output.

Earlier works use hand-craft rules or statistical methods to solve DST tasks. While widely used in industry dialogue systems, rule-based DST methods (Goddeau et al. 1996) have many restrictions such as limited generalization, high error rate, low domain adaptability, etc (Williams 2014). Statistical methods (Lee 2013; Lee and Eskenazi 2013; Ren et al. 2013; Williams 2013, 2014) also suffer from noisy conditions and ambiguity (Young et al. 2010).

Recently, many neural trackers have emerged. Neural trackers have multiple advantages over rule-based and statistical trackers. In general, they are categorized into two streams. The first stream has predefined slot names and values, and each turn the DST module tries to find the most appropriate slot-value pairs based on the dialogue history; the second stream does not have a fixed slot value list, so the DST module tries to find the values directly from the dialogue context or generate values based on the dialogue context. Obviously, the latter one is more flexible and in fact, more and more works are solving DST in a second way. We discuss the works of both categories here.

Neural trackers with predefined slot names and values The first stream can be viewed as a multi-class or multi-hop classification task. For multi-class classification DST, the tracker predicts the correct class from multiple values but this method suffers from high complexity when the value set grows large. On the other hand, for the multi-hop classification tasks, the tracker reads only one slot-value pair at a time and performs binary prediction. Working in this fashion reduces the model complexity but raises the system reaction time since for each slot there will be multiple tracking processes. Henderson et al. (2013) was the first who used a deep learning model in the DST tasks. They integrated many feature functions (e.g., SLU score, Rank score, Affirm score, etc.) as the input of a neural network, then predict the probability of each slot-value pair. Mrkšić et al. (2015) applied an RNN as a neural tracker to gain awareness of dialogue context. Mrkšić et al. (2017) proposed a multi-hop neural tracker which took the system output and user utterances as the first two inputs (to model the dialogue context), and the candidate slot-value pairs as the third input. The tracker finally made a binary prediction on the current slot-value pair based on the dialogue history.

Neural trackers with unfixed slot names and values The second stream attracts more attention because it not only reduces the model and time complexity of DST tasks but also facilitates end-to-end training of task-oriented dialogue systems. Moreover, it is also flexible when the target domain changes. Lei et al. (2018) proposed belief span, a text span of the dialogue context corresponding to a specific slot. They built a two-stage CopyNet to copy and store slot values from the dialogue history. The slots were stored to prepare for neural response generation. The belief span facilitated the end-to-end training of dialogue systems and increased the tracking accuracy in out-of-vocabulary cases. Based on this, Lin et al. (2020c) proposed the minimal belief span and argued that it was not scalable to generate belief states from scratch when the system interacted with APIs from diverse domains. The proposed MinTL framework operated *insertion (INS)*, *deletion (DEL)* and *substitution (SUB)* on the dialogue state of last turn based on the context and the minimal belief span. Wu et al. (2019a) proposed the TRADE model. The model also applied the copy mechanism and used a soft-gated pointer-generator to generate the slot value based on the domain-slot pair and encoded dialogue context. Quan and Xiong (2020) argued that simply concatenating the dialogue context was not preferable. Alternatively, they used *[sys]* and *[usr]* to discriminate the system and user messages. This simple long context

modeling method achieved a 7.03% improvement compared with the baseline. Cheng et al. (2020) proposed Tree Encoder-Decoder (TED) architecture which utilized a hierarchical tree structure to represent the dialogue states and system acts. The TED generated tree-structured dialogue states of the current turn based on the dialogue history, dialogue action, and dialogue state of the last turn. This approach led to a 20% improvement in the state-of-the-art DST baselines which represented dialogue states and user goals in a flat space. Chen et al. (2020a) built an interactive encoder to exploit the dependencies within a turn and between turns. Furthermore, they used the attention mechanism to construct the slot-level context for user and system respectively, which were embedding vectors based on which the generator copied values from the dialogue context. Shan et al. (2020) applied BERT to perform multi-task learning and generated the dialogue state. They first encoded word-level and turn-level contexts. Then they retrieved the relevant information for each slot from the context by applying both word-level and turn-level attention. Furthermore, the slot values were predicted based on the retrieved information. Similarly, Wang et al. (2020e) used BERT for slot value prediction. They performed Slot Attention (SA) to retrieve related spans and Value Normalization (VN) to convert the spans into final values. Huang et al. (2020c) proposed Meta-Reinforced MultiDomain State Generator (MERET), which was a dialogue state generator further finetuned with policy gradient reinforcement learning.

3.3 Policy learning

The Policy learning module is the other module of a dialogue manager. This module controls which action will be taken by the system based on the output dialogue states from the DST module. Assuming that we have the dialogue state S_t of the current turn and the action set $A = \{a_1, \ldots, a_n\}$, the task of this module is to learn a mapping function f: $S_t \rightarrow a_i \in A$. This module is comparatively simpler than other modules in terms of task definition but actually, the task itself is challenging (Peng et al. 2017). For example, in the tasks of movie ticket and restaurant table booking, if the user books a two-hour movie slot and intends to go for dinner after that, then the agent should be aware that the time gap between the movie slot and restaurant slot has to be more than two hours since the commuting time from the cinema to the restaurant should be considered.

Supervised learning and reinforcement learning are mainstream training methods for dialogue policy learning (Chen et al. 2017a). Policies learned in a supervised fashion exhibit great decision-making ability (Su et al. 2016; Dhingra et al. 2017; Williams et al. 2017; Liu and Lane 2017). In some specific tasks, the supervised policy model can complete tasks precisely, but the training process totally depends on the quality of training data. Moreover, the annotated datasets require intensive human labor, and the decision ability is restricted by the specific task and domain, showing weak transferring capability. With the prevalence of reinforcement learning methods, more and more task-oriented dialogue systems use reinforcement learning to learn the policy. The dialogue policy learning fits the reinforcement learning setting since the agent of reinforcement learning learns a policy to map environment states to actions as well.

Usually, the environment of reinforce policy learning is a user or a simulated user in which setting the training is called online learning. However, it is data- and time-consuming to learn a policy from scratch in the online learning scenario, so the warm-start method is needed to speed up the training process. Henderson et al. (2008) used expert data to

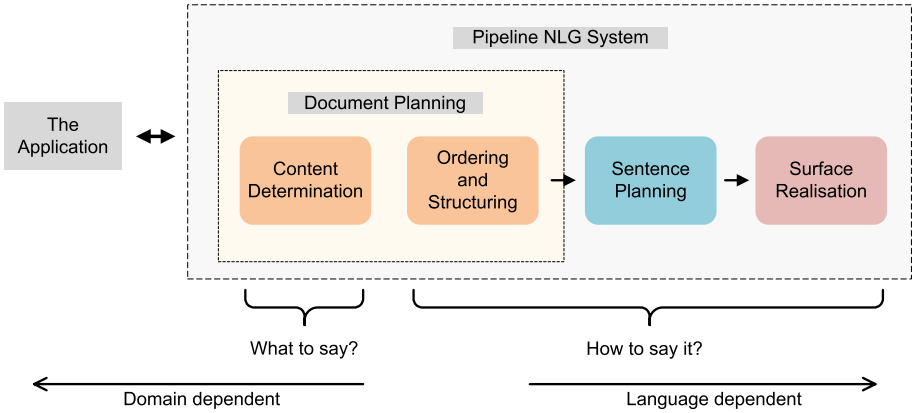


Fig. 15 The pipeline NLG system

restrict the initial action space exploration. Chen et al. (2017b) applied a teacher-student learning framework to transfer the teacher's expert knowledge to the target network in order to warm-start the system.

Reinforcement policy learning techniques Almost all recent dialogue policy learning works are based on reinforcement learning methods. Online learning is an ideal approach to get training samples iteratively for a reinforcement learning agent, but human labor is very limited. Zhang et al. (2019d) proposed Budget-Conscious Scheduling (BCS) to better utilize limited user interactions, where the user interaction is seen as the budget. The BCS used a probability scheduler to allocate the budget during training. Also, a controller decided whether to use real user interactions or simulated ones. Furthermore, a goal-based sampling model was applied to simulate the experiences for policy learning. Such a budget-controlling mechanism achieved ideal performance in the practical training process. Considering the difficulty of getting real online user interactions and the huge amount of annotated data required for training user simulators, Takanobu et al. (2020) proposed Multi-Agent Dialog Policy Learning, where they have two agents interacting with each other, performing both user and agent, learning the policy simultaneously. Furthermore, they incorporated a role-specific reward to facilitate role-based response generation. A High task completion rate was observed in experiments. Wang et al. (2020d) introduced Monte Carlo Tree Search with Double-q Dueling network (MCTS-DDU), where decision-time planning was proposed instead of background planning. They used the Monte Carlo simulation to perform a tree search of the dialogue states. Gordon-Hall et al. (2020) trained expert demonstrators in a weakly supervised fashion to perform Deep Q-learning from Demonstrations (DQfD). Furthermore, Reinforced Fine-tune Learning was proposed to facilitate domain transfer. In reinforce dialogue policy learning, the agent usually receives feedback at the end of the dialogue, which is not efficient for learning. Huang et al. (2020b) proposed an innovative reward learning method that constrains the dialogue progress according to the expert demonstration. The expert demonstration could either be annotated or not, so the approach was not labor-intensive. Wang et al. (2020b) proposed to co-generate the dialogue actions and responses to maintain the inherent semantic structures of dialogue. Similarly, Le et al. (2020b) proposed a unified framework to simultaneously perform

dialogue state tracking, dialogue policy learning, and response generation. Experiments showed that unified frameworks have a better performance both in their sub-tasks and in their domain adaptability. Xu et al. (2020a) used a knowledge graph to provide prior knowledge of the action set and solved policy learning tasks in a graph-grounded fashion. By combining a knowledge graph, a long-term reward was obtained to provide the policy agent with a long-term vision while choosing actions. Also, the candidate actions were of higher quality due to prior knowledge. The policy learning was further performed in a more controllable way.

3.4 Natural language generation

Natural Language Generation (NLG) is the last module of a task-oriented dialogue system pipeline. It manages to convert the dialogue actions generated from the dialogue manager into a final natural language representation. E.g., Assuming "*Inform (name = Wonder Woman; genre = Action; desti = Golden Village)*" to be the dialogue action from the policy learning module, then the NLG module converts it into language representations such as "*There is an action movie named Wonder Woman at Golden Village.*"

Traditional NLG modules are pipeline systems. Defined by Siddharthan (2001), the standard pipeline of NLG consists of four components, as shown in Fig. 15.

The core modules of this pipeline are Content Determination, Sentence Planning, and Surface Realization, as proposed by Reiter (1994). Cahill et al. (1999) further improved the NLG pipeline by adding three more components: lexicalization, referring expression generation, and aggregation. However, this model has a drawback in that the input of the system is ambiguous.

End-to-end NLG techniques Deep learning methods were further applied to enhance the NLG performance and the pipeline is collapsed into a single module. End-to-end natural language generation has achieved promising improvements and is the most popular way to perform NLG in recent work. Wen et al. (2015a) argued that language generation should be fully data-driven and not depend on any expert rules. They proposed a statistical language model based on RNNs to learn response generation with semantic constraints and grammar trees. Additionally, they used a CNN reranker to further select better responses. Similarly, an LSTM model was used by Wen et al. (2015b) to learn sentence planning and surface realization simultaneously. Tran and Nguyen (2017) further improved the generation quality on multiple domains using GRU. The proposed generator consistently generated high-quality responses on multiple domains. To improve the domain adaptability of recurrent models, Wen et al. (2016b) proposed to first train the recurrent language model on data synthesized from out-of-domain datasets, then finetune on a comparatively smaller in-domain dataset. This training strategy was proved effective in human evaluation. Context-awareness is important in dialogue response generation because only depending on the dialogue action of the current turn may cause illogical responses. Zhou et al. (2016) built an attention-based Context-Aware LSTM (CA-LSTM) combining target user questions, all semantic values, and dialogue actions as input to generate context-aware responses in QA. Likewise, Dušek and Jurčíček (2016a) concatenated the preceding user utterance with the dialogue action vector and fed it into an LSTM model. Dušek and Jurčíček (2016b) put a syntax constraint upon their neural response generator. A two-stage sequence generation process was proposed. First, a syntax dependency tree was generated to have a structured representation of the dialogue utterance. The generator in the second stage integrated sentence planning and surface realization and produced natural language representations.

Robust natural language generation More recent works have focused on the reliability and quality of generated responses. A tree-structured semantic representation was proposed by Balakrishnan et al. (2019) to achieve better content planning and surface realization performance. They further designed a novel beam search algorithm to improve the semantic correctness of the generated response. To avoid mistakes such as slot value missing or redundancy in generated responses, Li et al. (2020e) proposed Iterative Rectification Network (IRN), a framework trained with supervised learning and finetuned with reinforcement learning. It iteratively rectified generated tokens by incorporating slot inconsistency penalty into its reward. Golovanov et al. (2019) applied large-scale pretrained models for NLG tasks. After comparing single-input and multi-input methods, they concluded that different types of input contexts will cause different inductive biases in generated responses and further proposed to utilize this characteristic to better adapt a pretrained model to a new task. Baheti et al. (2020) solved NLG reliability problem in conversational QA. Though with different pipeline structures, they used similar methods to increase the fluency and semantic correctness of the generated response. They proposed Syntactic Transformations (STs) to generate candidate responses and used a BERT to rank their qualities. These generated responses can be viewed as an augmentation of the original dataset to be further used in NLG model learning. Oraby et al. (2019) proposed a method to create datasets with rich style markups from easily available user reviews. They further trained multiple NLG models based on generated data to perform joint control of semantic correctness and language style. Similarly, Elder et al. (2020) put forward a data augmentation approach that put a restriction on response generation. Though this restriction caused dull and less diverse responses, they argued that in task-oriented systems, reliability was more important than diversity.

3.5 End-to-end methods

The modules discussed above can achieve good performance in their respective tasks, with the help of recent relevant advances. However, there exist two significant drawbacks in modular systems (Zhao and Eskenazi 2016): (1) Modules in many pipeline systems are sometimes not differentiable, which means that errors from the end are not able to be propagated back to each module. In real dialogue systems training, usually, the only signal is the user response, while other supervised signals like dialogue states and dialogue actions are scarce. (2) Though the modules jointly contribute to the success of a dialogue system, the improvement of one module may not necessarily raise the response accuracy or quality of the whole system. This causes additional training for other modules, which is labor-intensive and time-consuming. Additionally, due to the handcrafted features in pipeline task-oriented systems such as dialogue states, it is usually hard to transfer modular systems to another domain, since the predefined ontologies require modification.

There exist two main methods for the end-to-end training of task-oriented dialogue systems. One is to make each module of a pipeline system differentiable, then the whole pipeline can be viewed as a large differentiable system and the parameters can be optimized by back-propagation in an end-to-end fashion (Le et al. 2020b). Another way is to use only one end-to-end module to perform both knowledge base retrieval and response generation, which is usually a multi-task learning neural model.

End-to-end trainable pipeline TOD The increasing applications of neural models have made it possible for modules to be differentiable. While many modules are easily differentiable, there remains one task that makes differentiation challenging: the knowledge

base query. Many task-oriented dialogue systems require an external knowledge source to retrieve related knowledge facts required by the user. For example, in the restaurant table booking task, the knowledge fact can be an available slot of one specific restaurant. Traditional methods use a symbolic query to match entries based on their attributes. The system performs semantic parsing on the user message to represent a symbolic query according to the user goal (Li et al. 2017d; Williams and Zweig 2016; Wen et al. 2017). However, this retrieval process is not differentiable, which prevents the whole framework from being end-to-end trainable. With the application of key-value memory networks (Miller et al. 2016), Eric et al. (2017) used the key-value retrieval mechanism to retrieve relevant facts. The proposed architecture was augmented with the attention mechanism to compute the relevance between utterance representations of dialogue and key representations of the knowledge base. Dhingra et al. (2017) presented a soft retrieval mechanism that uses a "soft" posterior distribution over the knowledge base to replace the symbolic queries. They further combined this soft retrieval mechanism into a reinforcement learning framework to achieve complete end-to-end training based on user feedback. Williams et al. (2017) proposed Hybrid Code Networks (HCNs), which encoded domain-specific knowledge into software and system action templates, achieving the differentiability of the knowledge retrieval module. They did not explicitly model the dialogue states but instead learned the latent representation and optimized the HCN using supervised learning and reinforcement learning jointly. Ham et al. (2020) used GPT-2 to form a neural pipeline and perform domain prediction, dialogue state tracking, policy learning, knowledge retrieval, and response generation in a pipeline fashion. The system could easily interact with external systems because it outputs explicit intermediate results from each module and thus is interpretable. Likewise, Hosseini-Asl et al. (2020) built a neural pipeline with GPT-2 and explicitly generated results for each neural module as well.

End-to-end trainable single module TOD More recent works tend not to build their end-to-end systems in a pipeline fashion. Instead, they use complex neural models to implicitly represent the key functions and integrate the modules into one. Research in task-oriented end-to-end neural models focuses either on training methods or model architecture, which are the keys to response correctness and quality. Wang et al. (2019a) proposed an incremental learning framework to train their end-to-end task-oriented system. The main idea is to build an uncertainty estimation module to evaluate the confidence of appropriate responses generated. If the confidence score was higher than a threshold, then the response would be accepted, while a human response would be introduced if the confidence score was low. The agent could also learn from human responses using online learning. Dai et al. (2020) used model-agnostic meta-learning (MAML) to improve the adaptability and reliability jointly with only a handful of training samples in a real-life online service task. Similarly, Qian and Yu (2019) also trained the end-to-end neural model using MAML to facilitate the domain adaptation, which enables the model to first train on rich-resource tasks and then on new tasks with limited data. Lin et al. (2020c) proposed Minimalist Transfer Learning (MinTL) to plug-and-play large-scale pretrained models for domain transfer in dialogue task completion. To maintain the sequential correctness of generated responses, Wu et al. (2019b) trained an inconsistent order detection module in an unsupervised fashion. This module detected whether an utterance pair is ordered or not to guide the task-completing agent toward generating more coherent responses. He et al. (2020a) proposed a "Two-Teacher One-Student" training framework. In the first stage, the two teacher models were trained in a reinforcement learning framework, with the objective of retrieving knowledge facts and generating human-like responses respectively. Then in the second stage, the student network was forced to mimic the output of the teacher networks.

Thus, the expert knowledge of the two teacher networks was transferred to the student network. Balakrishnan et al. (2019) introduced a constrained decoding method to improve the semantic correctness of the responses generated by the proposed end-to-end system. Many end-to-end task-oriented systems used a memory module to store relevant knowledge facts and dialogue history. Chen et al. (2019b) argued that a single memory module was not enough for precise retrieval. They used two long-term memory modules to store the knowledge tuples and dialogue history respectively, and then a working memory was applied to control the token generation. Zhang et al. (2020b) proposed LAtent BElief State (LABES) model, which treated the dialogue states as discrete latent variables to reduce the reliance on turn-level DST labels. To solve the data insufficiency problem in some tasks, Gao et al. (2020a) augmented the response generation model with a paraphrasing model in their end-to-end system. The paraphrase model was jointly trained with the whole framework and it aimed to augment the training samples. Yang et al. (2020) leveraged the graph structure information of both a knowledge graph and the dialogue context-dependency tree. They proposed a recurrent cell architecture to learn representations on the graph and performed multi-hop reasoning to exploit the entity links in the knowledge graph. With the augmentation of graph information, consistent improvement was achieved on two task-oriented datasets.

3.6 Research challenges and hot topics

In this section, we review recent works in task-oriented dialogue systems and point out the frequently studied topics to provide some important research directions. This section can be seen as an augmentation of the literature review in previous sections discussing techniques developed for each module and focusing more on some specific problems to be solved in the current research community.

3.6.1 Pretrained models for NLU

The Natural Language Understanding task converts the user message into a predefined format of semantic slots. A popular way to perform NLU is by finetuning large-scale pretrained language models. Wu and Xiong (2020) compared many pretrained language models including BERT-based and GPT-based systems in three subtasks of task-oriented dialogue systems - domain identification, intent detection, and slot tagging. This empirical paper is aimed to provide insights and guidelines in pretrained model selection and application for related research. Wu et al. (2020a) pretrained TOD-BERT and outperformed strong baselines in the intent detection task. The model proposed also had a strong few-shot learning ability to alleviate the data insufficiency problem. Coope et al. (2020) proposed Span-ConveRT, which was a pretrained model designed for slot filling tasks. It viewed the slot filling task as a turn-based span extraction problem and also performed well in the few-shot learning scenario.

3.6.2 Domain transfer for NLU

Another challenge or hot topic in NLU research is the domain transfer problem, which is also the key issue of task-oriented dialogue systems. Hakkani-Tür et al. (2016) built an RNN-LSTM architecture for multitask learning of domain classification, intent detection,

and slot-filling problem. Training samples from multiple domains were combined in a single model where respective domain data reinforces each other. Bapna et al. (2017) used a multi-task learning framework to leverage slot name encoding and slot description encoding, thus implicitly aligning the slot-filling model across domains. Likewise, Lee and Jha (2019) also applied slot description to exploit the similar semantic concepts between slots of different domains, which solved the sub-optimal concept alignment and long training time problems encountered in past works involving multi-domain slot-filling.

3.6.3 Domain transfer for DST

Domain adaptability is also a significant topic for dialogue state trackers. The domain transfer in DST is challenging due to three main reasons (Ren et al. 2018): (1) Slot values in ontologies are different when the domain changes, which accounts for the incompatibility of models. (2) When the domain changes, the slot number will also change, causing different numbers of model parameters. (3) Hand-crafted lexicons make it difficult for generalization over domains. Mrkšić et al. (2015) used delexicalized n-gram features to solve the domain incompatibility problem by replacing all specified slot names and values with generic symbols. Lin et al. (2020c) introduced Levenshtein belief spans (Lev), which were short context spans relating to the user message. Different from previous methods which generated dialogue states from scratch, they performed substitution (SUB), deletion (DEL), and insertion (INS) based on past states to alleviate the dependency on annotated in-domain training samples. Huang et al. (2020c) applied model-agnostic meta-learning (MAML) to first learn on several source domains and then adapt on the target domain, while Campagna et al. (2020) improved the zero-shot transfer learning by synthesizing in-domain data using an abstract conversation model and the domain ontology. Ouyang et al. (2020) modeled explicit slot connections to exploit the existing slots appearing in other domains. Thus, the tracker could copy slot values from the connected slots directly, alleviating the burden of reasoning and learning. Wang et al. (2020e) proposed Value Normalization (VN) to convert supporting dialogue spans into state values and could achieve high accuracy with only 30% available ontology.

3.6.4 Tracking efficiency for DST

Tracking efficiency is another hot topic in dialogue state tracking challenges. Usually, there are multiple states within a dialogue, so how to compute the slot values without any redundant steps becomes very significant when attempting to reduce the reaction time of a system. Kim et al. (2020b) argued that predicting the dialogue state from scratch at every turn was not efficient. They proposed to first predict the operations to be taken on each of the slots (i.e., Carryover, Delete, Dontcare, Update), and then perform respective operations as predicted. Ouyang et al. (2020) used a slot connection mechanism to directly copy slot values from the source slot, which reduced the expense of reasoning. Hu et al. (2020) and Wang et al. (2020e) proposed slot attention to calculate the relations between the slot and dialogue context, thus only focusing on the relevant slots at each turn.

3.6.5 Training environment for PL

The environment of the Policy Learning framework has been a long-existing problem. Li et al. (2017d) built a user simulator to model the user feedback as the reward signal of an

environment. They modeled a stack-like user agenda to iteratively change the user goal and thus shift the dialogue states. While using a user simulator for environment modeling seems to be promising for that it involves less human interaction, Zhang et al. (2019d) argued that training a user simulator required a large amount of annotated data. Takanobu et al. (2020) proposed Multi-Agent Dialog Policy Learning, where they have two agents interact with each other, performing both user and agent, learning policy simultaneously. Furthermore, they incorporated a role-specific reward to facilitate role-based response generation and here both agents also acted as the environment of the other one.

3.6.6 Response consistency for NLG

Response consistency in NLG is a challenging problem since it cannot be solved by simply augmenting the training samples. Instead, additional corrections or regulations should be designed. Wen et al. (2015b) proposed the Semantically Controlled LSTM (SC-LSTM) which used a semantic planning gate to control the retention or abandonment of dialogue actions thus ensuring the response consistency. Likewise, Tran and Nguyen (2017) also applied a gating mechanism to jointly perform sentence planning and surface realization where dialogue action features were gated before entering GRU cells. Li et al. (2020e) proposed Iterative Rectification Network (IRN), which combined a slot inconsistency reward into the reinforcement learning framework. Thus, the model iteratively checked the correctness of slots and corresponding values.

3.6.7 End-to-end task-oriented dialogue systems

End-to-end systems are usually fully data-driven, which contributes to their robust and natural responses. However, because of the finiteness of annotated training samples, a hot research topic is figuring out how to increase the response quality of end-to-end task-oriented dialogue systems with limited data. Using rule-based methods to constrain response generation is a way to improve response quality. Balakrishnan et al. (2019) used linearized tree-structured representation as input to obtain control over discourse-level and sentence-level semantic concepts. Kale and Rastogi (2020) used templates to improve the semantic correctness of generated responses. They broke down the response generation into a two-stage process: first generating semantically correct but possibly incoherent responses based on the slots, with the constraint of templates; then in the second stage, pretrained language models were applied to re-organize the generated utterances into coherent ones. Training the network with reinforcement learning was another strategy to alleviate the reliance on annotated data. He et al. (2020a) trained two teacher networks using a reinforcement learning framework with the objectives of knowledge retrieval and response generation respectively. Then the student network learns to produce responses by mimicking the output of teacher networks. Training the network in a supervised way, Dai et al. (2020) alternatively tried to optimize the learning strategy to improve the learning efficiency of models given limited data. They combined the meta-learning algorithm with human-machine interaction and achieved significant improvement compared with strong baselines not trained with the meta-learning algorithms. A more direct way to solve the data finiteness problem in supervised learning was augmenting the dataset (Elder et al. 2020), which also improved the response quality to some extent. Additionally, pretraining large-scale models on a common corpus and then applying them in a domain that lacks annotated data is a popular approach in recent years (Henderson et al. 2019b; Mehri et al. 2019; Bao et al. 2020).

3.6.8 Retrieval methods for task-oriented dialogue systems

Retrieval-based methods are rare in task-oriented systems for the insufficiency of candidate entries to cover all possible responses which usually involve specific knowledge from external knowledge-base. However, Henderson et al. (2019b) argued that in some situations not relating to specific knowledge facts, retrieval-based methods were more precise and effective. They first pretrained the response selection model on general domain corpora and then finetuned on small target domain data. Experiments on six datasets from different domains proved the effectiveness of the pretrained response selection model. Lu et al. (2019b) constructed Spatio-temporal context features to facilitate response selection, and achieved significant improvements on the Ubuntu IRC dataset.

4 Open-domain dialogue systems

This section discusses open-domain dialogue systems, which are also called chit-chat dialogue systems or non-task-oriented dialogue systems. Almost all state-of-the-art open-domain dialogue systems are based on neural methods. We organize this section by first briefly introducing the concepts of different branches of open-domain dialogue systems, and then we focus on different research challenges and hot topics. We view these challenges and hot topics as different research directions in open-domain dialogue systems.

Instead of managing to complete tasks, open-domain dialogue systems aim to perform chit-chat with users without the task and domain restriction (Ritter et al. 2011) and are usually fully data-driven. Open-domain dialogue systems are generally divided into three categories: generative systems, retrieval-based systems, and ensemble systems. Generative systems apply sequence-to-sequence models to map the user message and dialogue history into a response sequence that may not appear in the training corpus. By contrast, retrieval-based systems try to find a pre-existing response from a certain response set. Ensemble systems combine generative methods and retrieval-based methods in two ways: retrieved responses can be compared with generated responses to choose the best among them; generative models can also be used to refine the retrieved responses (Zhu et al. 2019; Song et al. 2016; Qiu et al. 2017; Serban et al. 2017a). Generative systems can produce flexible and dialogue context-related responses while sometimes they lack coherence and tend to make dull responses. Retrieval-based systems select responses from human response sets and thus are able to achieve better coherence in surface-level language. However, retrieval systems are restricted by the finiteness of the response sets and sometimes the responses retrieved show a weak correlation with the dialogue context (Zhu et al. 2019).

In the next few subsections, we discuss some research challenges and hot topics in open-domain dialogue systems. We aim to to help researchers quickly grasp the current research trends via a systematic discussion on articles solving certain problems.

4.1 Context awareness

Dialogue context consists of user and system messages and is an important source of information for dialogue agents to generate responses because dialogue context decides the conversation topic and user goal (Serban et al. 2017b). A context-aware dialogue agent responds not only depending on the current message but also based on the conversation history. The earlier deep learning-based systems added up all word representations in dialogue

history or used a fixed-size window to focus on the recent context (Sordoni et al. 2015b; Li et al. 2016a). Serban et al. (2016) proposed Hierarchical Recurrent Encoder-Decoder (HRED), which was ground-breaking in building context-awareness dialogue systems. They built a word-level encoder to encode utterances and a turn-level encoder to further summarize and deliver the topic information over past turns. Xing et al. (2018) augmented the hierarchical neural networks with the attention mechanism to help the model focus on more meaningful parts of dialogue history.

Both generative and retrieval-based systems rely heavily on dialogue context modeling. Shen et al. (2019) proposed Conversational Semantic Relationship RNN (CSRR) to model the dialogue context in three levels: utterance-level, pair-level, and discourse-level, capturing content information, user-system topic, and global topic respectively. Zhang et al. (2019c) argued that the hierarchical encoder-decoder does not lay enough emphasis on certain parts when the decoder interacted with dialogue contexts. Also, they claimed that attention-based HRED models also suffered from position bias and relevance assumption insufficiency problems. Therefore, they proposed ReCoSa, whose architecture was inspired by the transformer. The model first used a word-level LSTM to encode dialogue contexts, and then self-attention was applied to update the utterance representations. In the final stage, an encoder-decoder attention was computed to facilitate the response generation process. Additionally, Mehri et al. (2019) examined several applications of large-scale pretrained models in dialogue context learning, providing guidance for large-scale network selection in context modeling.

Some works propose structured attention to improve context-awareness. Qiu et al. (2020) learned structured dialogue context by combining structured attention with a Variational Recurrent Neural Network (VRNN). Comparatively, Ferracane et al. (2019) examined the RST discourse tree model proposed by Liu and Lapata (2018) and observed little or even no discourse structures in the learned latent tree. Thus, they argued that structured attention did not benefit dialogue modeling and sometimes might even harm the performance.

Interestingly, Feng et al. (2020b) not only utilized dialogue history, but also future conversations. Considering that in real inference situations dialogue agents cannot be explicitly aware of future information, they first trained a scenario-based model jointly on past and future context and then used an imitation framework to transfer the scenario knowledge to a target network.

Better context modeling improves the response selection performance in retrieval-based dialogue systems (Jia et al. 2020). Tao et al. (2019) proposed Interaction-over-Interaction network (IoI), which consisted of multiple interaction blocks to perform deeper interactions between dialogue context and candidate responses. Jia et al. (2020) organized the dialogue history into conversation threads by performing classifications on their dependency relations. They further used a pretrained Transformer model to encode the threads and candidate responses to compute the matching score. Lin et al. (2020b) argued that response-retrieval datasets should not only be annotated with relevant or irrelevant responses. Instead, a greyscale metric should be used to measure the relevance degree of a response given the dialogue context, thus increasing the context-awareness ability of retrieval models.

Dialogue rewriting problem aims to convert several messages into a single message conveying the same information and dialogue context awareness is very crucial to this task (Xu et al. 2020b). Su et al. (2019a) modeled multi-turn dialogues via dialogue rewriting and benefited from the conciseness of rewritten utterances.

4.2 Response coherence

Coherence is one of the qualities that a good generator seeks (Stent et al. 2005). Coherence means maintaining logic and consistency in a dialogue, which is essential in an interaction process for that a response with weak consistency in logic and grammar is hard to understand. Coherence is a hot topic in generative systems but not in retrieval-based systems because candidate responses in retrieval methods are usually human responses, which are naturally coherent.

Refining the order or granularity of sentence functions is a popular strategy for improving the language coherence. Wu et al. (2019b) improved the response coherence via the task of inconsistent order detection. The dialogue systems learned response generation and order detection jointly, which was self-supervised multi-task learning. Xu et al. (2019) presented the concept of meta-words. Meta-words were diverse attributes describing the response. Learning dialogue based on meta-words helped promote response generation in a more controllable way. Liu et al. (2019) used three granularities of encoders to encode raw words, low-level clusters, and high-level clusters. The architecture was called Vocabulary Pyramid Network (VPN), which performed a multi-pass encoding and decoding process on hierarchical vocabularies to generate coherent responses. Shen et al. (2019) also built a three-level hierarchical dialogue model to capture richer features and improved the response quality. Ji et al. (2020) built Cross Copy Networks (CCN), which used a copy mechanism to copy from similar dialogues based on the current dialogue context. Thus, the system benefited from the pre-existing coherent responses, which alleviated the need of performing the reasoning process from scratch.

Many work employ strategies to achieve response coherence on a higher level, which improves the overall quality of the generated responses. Li et al. (2020c) improved the logical consistency of generated utterances by incorporating an unlikelihood loss to control the distribution mismatches. Bao et al. (2019) proposed a Generation-Evaluation framework that evaluated the qualities, including coherence, of the generated response. The feedback was further seen as a reward signal in the reinforcement learning framework and guided to a better dialogue strategy via policy gradient, thus improving the response quality. Gao et al. (2020b) raised response quality by ranking generated responses based on user feedbacks like upvotes, downvotes, and comments on social networks. Zhu et al. (2019) built a retrieval-enhanced generation model, which enhanced the generated responses in two ways. First, a discriminator was trained with the help of a retrieval system, and then the generator was trained in a GAN framework under the supervision signal of a discriminator. Second, retrieved responses were also used as a part of the generator input to provide a coherent example for the generator. Xu et al. (2020a) achieved a global coherent dialogue by constructing a knowledge graph from corpora. They further performed graph walks to decide “what to say" and “how to say", thus improving the dialogue flow coherence. Mesgar et al. (2020) proposed an assessment approach for dialogue coherence evaluation by combining the dialogue act prediction in a multi-task learning framework and learned rich dialogue representations.

There also evolve some data-wise methods for better response coherence. Bi et al. (2019) proposed to annotate sentence functions in existing conversation datasets to improve the sentence logic and coherence of generated responses. Akama et al. (2020) focused on data effectiveness as well. They filtered out low-quality utterance pairs by scoring the relatedness and connectivity, which was proved to be effective in improving the response

coherence. Akama et al. (2020) presented a method for evaluating dataset utterance pairs' quality in terms of connectedness and relatedness. The proposed scoring technique is based on research findings that have been widely disseminated in the conversation and linguistics communities. Lison and Bibauw (2017) included a weighting model in their neural architecture. The weighting model, which is based on conversation data, assigns a numerical weight to each training sample that reflects its intrinsic quality for dialogue modeling and achieved good result in experiments.

4.3 Response diversity

The Bland and generic response is a long-existing problem in generative dialogue systems. Because of the high frequency of generic responses like *I don't know* in training samples and the beam search decoding scheme of neural sequence-to-sequence models, generative dialogue systems tend to respond with universally acceptable but meaningless utterances (Serban et al. 2016; Vinyals and Le 2015; Sordoni et al. 2015b). For example, to respond to the user message *I really want to have a meal* the agent tends to choose simple responses like *It's OK* instead of responding with more complicated sentences like recommendations and suggestions.

Earlier works solve this challenge by modifying the decoding objective or adding a reranking process. Li et al. (2016a) replaced the traditional likelihood objective $p(R|C)$ with mutual information. The optimization of the mutual information objective aims to achieve Maximum Mutual Information (MMI). Specifically, the task is to find the best response R based on the dialogue context C, in order to maximize their mutual information:

$$\begin{aligned}\hat{R} &= arg\max_{R} log\frac{P(C,R)}{P(C)P(R)} \\ &= arg\max_{R} logP(R|C) - logP(R)\end{aligned} \tag{49}$$

The objective $p(R|C)$ causes the model to choose responses with high probability even if the response is unconditionally frequent in the dataset, thus causing it to ignore the content of C. Maximizing the mutual information as Equation (49) solves this issue by achieving a trade-off between safety and relativity.

With a similar intuition as described above, increasing response diversity by modifying the decoding scheme at inference time has been explored in earlier works. Vijayakumar et al. (2016) combined a dissimilarity term into the beam search objective and proposed Diverse Beam Search (DBS) to promote diversity. Similarly, Shao et al. (2017) proposed a stochastic beam search algorithm by performing stochastic sampling when choosing top-B responses. In the beam search algorithm, siblings sharing the same parent nodes tended to guide to similar sequences. Inspired by this, Li et al. (2016b) penalized siblings sharing the same parent nodes using an additional term in the beam search objective. This encouraged the algorithm to search more diverse paths by expanding from different parent nodes. Some works further added a reranking stage to select more diverse responses in the generated N-best list (Li et al. 2016a; Sordoni et al. 2015b; Shao et al. 2017).

A user message can be mapped into multiple acceptable responses, which is also known as the one-to-many mapping problem. Qiu et al. (2019) considered the one-to-many mapping problem in open-domain dialogue systems and proposed a two-stage generation model to increase response diversity—the first stage extracting common features of multiple ground truth responses and the second stage extracting the distinctive ones. Ko et al.

(2020) solved the one-to-many mapping problem via a classification task to learn latent semantic representations. So that given one example response, different ones could be generated by exploring the semantically close vectors in the latent space.

Different training strategies have been proposed to increase response diversity. Bao et al. (2019) used human instinct or pre-defined objective as a reward signal in a reinforcement learning setting to prompt the agent to avoid generating dull responses. Still, in a reinforcement learning framework, Zhu et al. (2020) performed counterfactual reasoning to explore the potential response space. Given a pre-existing response, the model inferred another policy, which represented another possible response, thus increasing the response diversity. He and Glass (2020) used a negative training method to minimize the generation of bland responses. They first collected negative samples and then gave negative training signals based on these samples to fine-tune the model, impeding the model to generate bland responses. To achieve a better performance, Du and Black (2019) synthesized different dialogue models designed for response diversity based on boosting training. The ensemble model significantly outperformed each of its base models.

Utilizing external knowledge sources is another way to improve the diversity of generated responses because it can enrich the content. Wu et al. (2020b) built a common-sense dialogue generation model which seeks highly related knowledge facts based on the dialogue history. Likewise, Su et al. (2020a) incorporated external knowledge sources to diversify the response generation, but the difference was that they utilized non-conversational texts like news articles as relevant knowledge facts, which were obviously easier to obtain. Tian et al. (2019) used a memory module to abstract and store useful information in the training corpus for generating diverse responses.

Another approach to diversify the response generation is to make modifications to the training corpus. Csáky et al. (2019) solved the challenge by filtering out the generic responses in the dataset using an entropy-based algorithm, which was simple but effective. Augmented with human feedback data, Gao et al. (2020b) proposed that the generated responses could be reranked via a response ranking framework trained on the human feedback data and responses with higher quality including diversity were selected. Stasaski et al. (2020) proposed to change the data collection pipeline by iteratively computing the diversity of responses from different human participants in dataset construction and selected those participants who tend to generate informative and diverse responses.

4.4 Speaker consistency and personality-based response

In open-domain dialogue systems, one big issue is that the responses are entirely learned from training data. The inconsistent response may be received when asking the system about some personal facts (e.g., age, hobbies). If the dataset contains multiple utterance pairs about the query of age, then the response generated tends to be shifting, which is unacceptable because personal facts are usually not random. Thus, for a data-driven chatbot, it is necessary to be aware of its role and respond based on a fixed persona.

Explicitly modeling the persona is the main strategy in recent works. Liu et al. (2020b) proposed a persona-based dialogue generator consisting of a Receiver and a Transmitter. The receiver was responsible for modeling the interlocutor's persona through several turns' chat while Transmitter generated utterances based on the persona of agent and interlocutor, together with conversation content. The proposed model supported conversations between two persona-based chatbots by modeling each other's persona. Without training with additional Natural Language Inference labels, Kim et al. (2020a) built an imaginary listener

following a normal generator, which reasoned over the tokens generated by the generator and predicted a posterior distribution over the personas in a certain space. After that, a self-conscious speaker generated tokens aligned with the predicted persona. Likewise, Boyd et al. (2020) used an augmented GPT-2 to reason over the past conversations and model the target actor's persona, conditioning on which persona consistency was achieved.

Responding with personas needs to condition on some persona descriptions. For example, to build a generous agent, descriptions like "*I am a generous person*" are needed as a part of the model input. However, these descriptions require hand-crafted feature design, which is labor intensive. Madotto et al. (2019) proposed to use Model-Agnostic Meta-Learning (MAML) to adapt to new personas with only a few training samples and needed no persona description. Majumder et al. (2020a) relied on external knowledge sources to expand current persona descriptions so that richer persona descriptions were obtained, and the model could associate current descriptions with some commonsense facts.

Song et al. (2020a) argued that traditional persona-based systems were one-stage systems and the responses they generated still contain many persona inconsistent words. To tackle this issue, they proposed a three-stage architecture to ensure persona consistency. A generate-delete-rewrite mechanism was implemented to remove the unacceptable words generated in prototype responses and rewrite them.

4.5 Empathetic response

Empathy means being able to sense other people's feelings (Ma et al. 2020b; Li et al. 2022). An empathetic dialogue system can sense the user's emotional changes and produce appropriate responses with a certain sentiment (Tu et al. 2022). This is an essential topic in chit-chat systems because it directly affects the user's feeling and to some extent decides the response quality. Industry systems such as Microsoft's Cortana, Facebook M, Google Assistant, and Amazon's Alexa are all equipped with empathy modules (Wang et al. 2020g).

There are two ways to generate utterances with emotion: one is to use explicit sentiment words as a part of input; another is to implicitly combine neural words (Song et al. 2019). Song et al. (2019) proposed a unified framework that uses a lexicon-based attention to explicitly plugin emotional words and a sequence-level emotion classifier to classify the output sequence, implicitly guiding the generator to generate emotional responses through backpropagation. Zhong et al. (2020) used CoBERT for persona-based empathetic response selection and further investigated the impact of persona on empathetic responses. Smith et al. (2020) blended the skills of being knowledgeable, empathetic, and role-aware in one open-domain conversation model and overcame the bias issue when blending these skills.

Since the available datasets for empathetic conversations are scarce, Rashkin et al. (2019) provided a new benchmark and dataset for empathetic dialogue systems. Oraby et al. (2019) constructed a dialogue dataset with rich emotional markups from user reviews and further proposed a novel way to generate similar datasets with rich markups.

4.6 Controllable generation

Controllable dialogue generation is an important line of work in open-domain dialogue systems since solely learning from data sample distributions causes many uncertain responses. Some of the dialogue systems are grounded on some external knowledge such

as knowledge graph and documents. However, grounding alone without explicit control and semantic targeting may induce output that is accurate but vague.

We may get some inspirations from the prior work on language generation and machine translation since similarly to dialogue systems they are generation-based or seq-to-seq problems. Some related work aimed to enforce user-specified constraints, most notably using lexical constraints (Hokamp and Liu 2017; Hu et al. 2019; Miao et al. 2019). These methods exclusively use constraints at inference time. Constraints can be included into the latent space during training, resulting in better predictions. Other studies (See et al. 2019; Keskar et al. 2019; Tang et al. 2019) have looked at non-lexical constraints, but they haven't looked into how they can help with grounding external knowledge. These publications also assume that the system can always be given (gold) constraints, which limits the ability to demonstrate larger benefits of the approaches.

Controllable text generation has also been used to extract high-level style information from contextual information in text style transfer (Hu et al. 2017) and other tasks (Ficler and Goldberg 2017; Dong et al. 2017; Gao et al. 2019), allowing the former to be independently modified. Zhao et al. (2018) learns an interpretable representation for dialogue systems using discrete latent actions. While existing studies employ "style" descriptors (e.g., positive/negative, formal/informal) as control signals, Wu et al. (2020c) use specific lexical constraints to regulate creation, allowing for finer semantic control. Content planned generation (Wiseman et al. 2017; Hua and Wang 2019) focuses response generation on a small number of essential words or table entries. This line of work, on the other hand, does not require consideration of the discourse context, which is critical for response generation.

4.7 Conversation topic

Daily chats of people usually involve a topic or goal. Actually, a topic or goal is the key to keep each participant engaged in conversations and thus being essential to a chatbot. In real applications, a good topic model helps to retrieve related knowledge and guide the conversation instead of passively responding to the user's message (Xing et al. 2017). For example, if the user mentions "*I like sunny days*", a topic-aware system may reason over relevant external knowledge and produce responses like "*I know there is a nice park near the seaside, have you ever been there before?*". Thus, the agent pushes the conversation to a more engaging stage and enriches the dialogue content.

Almost all topic-aware dialogue agents need to model explicit topics, which can be entities from external knowledge-base, or topic embeddings that have some semantic meaning. Wu et al. (2019c) tried to change the traditional passive response fashion and radically pursue active guidance of conversation. The dialogue agent consists of a leader and a follower, where the leader reasons over a knowledge graph and decides the conversation topic. Likewise, a common-sense knowledge graph was used by Liu et al. (2020c) to lead the conversation topic and make recommendations. Tang et al. (2019) built a topic-aware retrieval-based chatbot. It aimed to guide the conversation topic to the target one step by step. It used a keyword predictor to predict turn-level keywords and selected the discourse-level keyword based on that. The discourse-level keyword was further fed into the retrieval model to retrieve responses regarding a certain topic. Chen and Yang (2020) built a multi-view sequence-to-sequence model to learn dialogue topics by first extracting dialogue structures of unstructured chit-chat dialogues, then generating topic summaries using BART decoder.

In some applications of certain scenarios the conversation topic is essential, and these are where the topic-aware dialogue agents can be applied to. Zhang and Danescu-Niculescu-Mizil (2020) studied the topic-aware chatbot in counseling conversations. In counseling conversations, the agent led the dialogue topic by deciding between empathetically addressing a situation within the current range and moving on to a new target resolution. Cao et al. (2019) studied chatbots in the psychotherapy treatment area and built a topic prediction model to forecast the behavior codes for upcoming conversations, thus guiding the dialogue.

4.8 Knowledge-grounded system

External knowledge such as common-sense knowledge is a significant source of information when organizing an utterance (Ni et al. 2022). Humans associate current conversation context with their experiences and memories and produce meaningful related responses, such capability results in the gap between human and machine chit-chat systems. As discussed, the earlier chit-chat systems are simply variants of machine translation systems, which can be viewed as sequence-to-sequence language models. However, dialogue generation is much more complicated than machine translation because of the higher freedom and vaguer constraints. Thus, chit-chat systems cannot simply consist of a sequence-to-sequence mapping since appropriate and informative responses are always related to some external common-sense knowledge. Instead, there must be a module incorporating world knowledge.

Many researchers devoted their research efforts to building knowledge-grounded dialogue systems. A representative model is memory networks introduced in Sect. 2.4. Knowledge grounded systems use Memory Networks to store external knowledge and the generator retrieves relevant knowledge facts from it at the generation stage (Ghazvininejad et al. 2018; Vougiouklis et al. 2016; Yin et al. 2016). Tian et al. (2019) built a memory-augmented conversation model. The proposed model abstracted from the training samples and stored useful ones in the memory module. Zhao et al. (2020b) built a knowledge-grounded dialogue generation system based on GPT-2. They combined a knowledge selection module into the language model and learned knowledge selection and response generation simultaneously. Lin et al. (2020a) proposed Knowledge-Interaction and knowledge Copy (KIC). They performed recurrent knowledge interactions during the decoding phase to compute an attention distribution over the memory. Then they performed knowledge copy using a knowledge-aware pointer network to copy knowledge words according to the attention distribution computed.

Documents contain large amount of knowledge facts, but they have a drawback that they are usually too long to retrieve useful information from (Li et al. 2019c). Li et al. (2019c) built a multi-turn document-grounded system. They used an incremental transformer to encode multi-turns' dialogue context and respective documents retrieved. In the generation phase, they designed a two-stage generation scheme. The first stage took dialogue context as input and generated coherent responses; the second stage utilized both the utterance from the first stage and the document retrieved for the current turn for response generation. In this case, selecting knowledge based on both dialogue context and generated response was called posterior knowledge selection, while selecting knowledge with only dialogue context was called prior knowledge selection, which only utilized prior information. Wang et al. (2020c) built a document quotation model

in online conversations and investigated the consistency between quoted sentences and latent dialogue topics.

Knowledge graph is another source of external information, which is becoming more and more popular in knowledge-grounded systems because of their structured nature. Jung et al. (2020) proposed a dialogue-conditioned graph traversal model for knowledge-grounded dialogue systems. The proposed model leveraged attention flows of two directions and fully made use of the structured information of knowledge graph to flexibly decide the expanding range of nodes and edges. Likewise, Zhang et al. (2020a) applied graph attention to traverse the concept space, which was a common-sense knowledge graph. The graph attention helped to move to more meaningful nodes conditioning on dialogue context. Xu et al. (2020a) applied knowledge graphs as an external source to control a coarse-level utterance generation. Thus, the conversation was supported by common-sense knowledge, and the agent guided the dialogue topic in a more reasonable way. Moon et al. (2019) built a retrieval system retrieving responses based on the graph reasoning task. They used a graph walker to traverse the graph conditioning on symbolic transitions of the dialogue context. Huang et al. (2020a) proposed Graph-enhanced Representations for Automatic Dialogue Evaluation (GRADE), a novel evaluation metric for open-domain dialogue systems. This metric considered both contextualized representations and topic-level graph representations. The main idea was to use an external knowledge graph to model the conversation logic flow as a part of the evaluation criteria.

Knowledge-grounded datasets containing context-knowledge-response triples are scarce and hard to obtain. Cho and May (2020) collected a large dataset consisting of more than 26000 turns of improvised dialogues which were further grounded with a larger movie corpus as external knowledge. Also tackling the data insufficiency problem, Li et al. (2020b) proposed a method that did not require context-knowledge-response triples for training and was thus data-efficient. They viewed knowledge as a latent variable to bridge the context and response. The variational approach learned the parameters of the generator from both a knowledge corpus and a dialogue corpus which were independent of each other.

4.9 Interactive training

Interactive training, also called human-in-loop training, is a unique training method for dialogue systems. Annotated data is fixed and limited, not being able to cover all dialogue settings. Also, it takes a long time to train a good system. But in some industrial products, the dialogue systems need not be perfect when accomplishing their tasks. Thus, interactive training is desirable because the dialogue systems can improve themselves via interactions with users anywhere and anytime, which is a more flexible and cheap way to finetune the parameters.

Training schemes with the above intuition have been developed in recent years. Li et al. (2017a) introduced a reinforcement learning-based online learning framework. The agent interacts with a human dialogue partner and the partner provides feedback as a reward signal. Asghar et al. (2017) first trained the agent with two-stage supervised learning, and then used an interaction-based reinforcement learning to finetune. Every time the user chose the best one from K responses generated by the pretrained model and then responded to this selected response. Instead of learning through being passively graded, Li et al. (2017b)

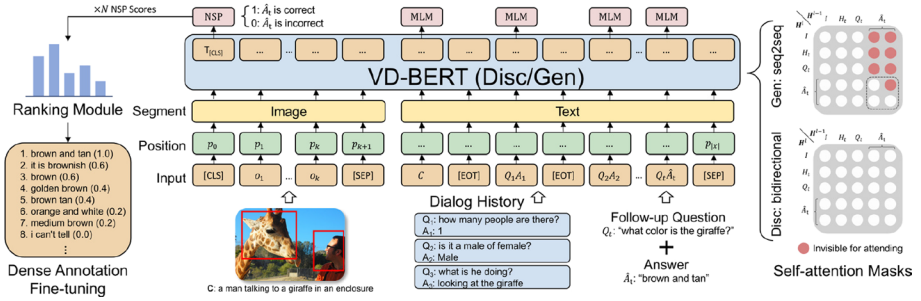


Fig. 16 The architecture of VD-BERT, a state-of-the-art visual dialogue system (Wang et al. 2020f)

proposed a model that actively asked questions to seek improvement. Active learning was applicable to both offline and online learning settings. Hancock et al. (2019) argued that most conversation samples an agent saw happened after it was pretrained and deployed. Thus, they proposed a framework to train the agent from the real conversations it participated in. The agent evaluated the satisfaction score of the user from the user's response to each turn and explicitly requested the user feedback when it thought that a mistake has been made. The user feedback was further used for learning. Bouchacourt and Baroni (2019) placed the interactive learning in a cooperative game and tried to learn a long-term implicit strategy via Reinforce algorithm. Some of these works have been adopted by industry products and are a very promising direction for study.

4.10 Visual dialogue

More and more researchers cast their eyes to a broader space and are not only restricted to NLP. The combination of CV and NLP giving rise to tasks like visual question answering attracted lots of interest. The VQA task is to answer a question based on the content of a picture or video. Recently, this has evolved into a more challenging task: visual dialogue, which conditions a dialogue on the visual information and dialogue history. The dialogue consists of a series of queries, and the query form is usually more informal, which is why it is more complicated than VQA.

Visual dialogue can be seen as a multi-step reasoning process over a series of questions (Gan et al. 2019). Gan et al. (2019) learned semantic representation of the question based on dialogue history and a given image, and recurrently updated the representation. Shuster et al. (2020c) proposed a set of image-based tasks and provided strong baselines. Wang et al. (2020f) employed R-CNN as an image encoder and fused the visual and dialogue modality with a VD-BERT. The proposed architecture achieved sufficient interactions between multi-turn dialogue and images. The proposed architecture is shown as an example model for Visual Dialogue tasks in Fig. 16.

Compared with image-grounded dialogue systems, video-grounded systems are more interesting but also more challenging. There are two main challenges of video dialogue, as claimed by Le and Hoi (2020). One is that both spatial and temporal features exist in the video, which increases the difficulty of feature extraction. Another is that video dialogue features span across multiple conversation turns and thus are more complicated. A GPT-2 model was applied by Le and Hoi (2020), being able to fuse multi-modality information over different levels. Likewise, Le et al. (2019) built a multi-modal transformer network to

A: Peaceful *B: Absentminded*	*A: Fearful* *B: Miserable*	*A: Erratic* *B: Skeptical*
A: I'm so thankful for this delicious food.	A: I just heard something out there and I have no idea what it was.	A: What is the difference between the forest and the trees? Oh look, dry pavement.
B: What is it called again?	B: It was probably a Wolf coming to eat us because you talk too much.	B: I doubt that's even a forest, it looks like a line of trees.
A: Not sure but fried goodness.	A: I would never go camping in the woods for this very reason.	A: There's probably more lame pavement on the other side!

Fig. 17 Three samples from the IMAGE-CHAT dataset (Shuster et al. 2020a)

incorporate information from different modalities and further applied a query-aware attention to extract context-related features from non-text modalities. Le et al. (2020a) proposed a Bi-directional Spatio-Temporal Learning (BiST) leveraging temporal-to-spatial and spatial-to-temporal reasoning process and could adapt to the dynamically evolving semantics in the video.

Some researchers hold different opinions on the effectiveness of dialogue history in visual dialogue. Takmaz et al. (2020) proposed that many expressions were already mentioned in previous turns and they built a visual dialogue model grounded on both image and conversation history. They further proved that better performance was achieved when grounding the model on dialogue context. However, Agarwal et al. (2020) argued that though with dialogue history the visual dialogue model could achieve better results, in fact only a small proportion of cases benefited from the history. Furthermore, they proved that existing evaluation metrics for visual dialogue promoted generic responses.

The visual dialogue task benefits a lot from the pretraining-based learning. The popularity of NLP pretraining sparked interest in multi-modal pretraining. VideoBERT (Sun et al. 2019b) is widely recognized as the pioneering work in the field of multimodal pretraining. It's a model that's been pre-trained on video frame features and text. CBT (Sun et al. 2019a), which is similarly pretrained on video-text pairs, is a contemporary work of VideoBERT. For video representation learning, Miech et al. (2020) used unlabeled narrated films. More researchers have focused their attention on visual-linguistic pretraining, inspired by the early work in multi-modal pretraining. For this objective, there are primarily two types of model designs. The single-stream model (Alberti et al. 2019; Chen et al. 2019c; Gan et al. 2020; Li et al. 2020a, 2019a, 2020d; Su et al. 2020b; Zhou et al. 2020b) is one example. (Li et al. 2020a) used a BERT model to process the concatenation of objects and words and pre-trained it with three standard tasks. Similar methods were proposed by Chen et al. (2019c) and Qi et al. (2020), but with more pretraining tasks and larger datasets. With an adversarial training technique, Gan et al. (2020) further enhanced the model. Su et al. (2020b) employed the same architecture, but incorporated single-modal data and pre-trained the object detector. Instead of using recognized objects, Huang et al. (2020d) sought to enter pixels directly. The object labels were used by Li et al. (2020d) to improve cross-modal alignment. Zhou et al. (2020b) suggested a single-stream model that learns both caption generation and VQA tasks at the same time. The two-stream model (Lu et al. 2019a, 2020; Tan and Bansal 2019; Yu et al. 2020) is another type of model architecture.

Tan and Bansal (2019) suggested a two-stream model with co-attention and solely used in-domain data to train the model. Lu et al. (2019a) introduced a similar architecture with a more complex co-attention model, which they pretrained with out-of-domain data, and Lu et al. (2020) improved VilBERT with multi-task learning. Yu et al. (2020) recently added the scene graph to the model, which improved performance. Aside from these studies, Singh et al. (2020) looked at the impact of pretraining dataset selection on downstream task performance.

The annotation of visual dialogue is laborious and thus the datasets are scarce. Recently, some researchers have tried to tackle the data insufficiency problem. Shuster et al. (2020a) collected a dataset (IMAGE-CHAT, shown in Figure 17) of image-grounded human-human conversations in which speakers are asked to perform role-playing based on an emotional mood or style offered, since the usage of such characteristics is also a significant factor in engagingness. Kamezawa et al. (2020) constructed a visual-grounded dialogue dataset. Interestingly, it additionally annotated the eye-gaze locations of the interlocutor in the image to provide information on what the interlocutor was paying attention to. Cogswell et al. (2020) proposed a method to utilize the VQA data when adapting to a new task, minimizing the requirement of dialogue data which is expensive to annotate.

5 Evaluation approaches

Evaluation is an essential part of research in dialogue systems. It is not only a way to assess the performance of agents, but it can also be a part of the learning framework which provides signals to facilitate the learning (Bao et al. 2019). This section discusses the evaluation methods in task-oriented and open-domain dialogue systems.

5.1 Evaluation methods for task-oriented dialogue systems

Task-oriented systems aim to accomplish tasks and thus have more direct metrics evaluating their performance such as task completion rate and task completion cost. Some evaluation methods also involve metrics like BLEU to compare system responses with human responses, which will be discussed later. In addition, human-based evaluation and user simulators are able to provide real conversation samples.

With regard to classification tasks such as intent detection and slot filling, traditional metrics such as accuracy, recall, precision, and F score (Powers 2020) can be used to evaluate their performance. These metrics are not only suitable for task-oriented dialogue systems, but also widely used in other domains such as Computer Vision, Graph, Recommendation Systems, etc.

Task Completion Rate is the rate of successful events in all task completion attempts. It measures the task completion ability of a dialogue system. For example, in movie ticket booking tasks, the Task Completion Rate is the fraction of dialogues that meet all requirements specified by the user, such as movie time, cinema location, movie genre, etc. The task completion rate was applied in many task-oriented dialogue systems (Walker et al. 1997; Williams 2007; Peng et al. 2017). Additionally, some works (Singh et al. 2002; Yih et al. 2015) used partial success rate.

Task Completion Cost is the resources required when completing a task. Time efficiency is a significant metric belonging to Task Completion Cost. In dialogue-related tasks, the

number of conversation turns is usually used to measure the time efficiency and dialogue with fewer turns is preferred when accomplishing the same task.

Human-based Evaluation provides user dialogues and user satisfaction scores for system evaluation. There are two main streams of human-based evaluation. One is to recruit human labor via crowdsourcing platforms to test and use a dialogue system. The crowdsource workers converse with the dialogue systems about predefined tasks and then metrics like Task Completion Rate and Task Completion Cost can be calculated. Another is computing the evaluation metrics in real user interactions, which means that evaluation is done after the system is deployed in real use.

User Simulator provides simulated user dialogues based on pre-defined rules or models. Since recruiting human labor is expensive and real user interactions are not available until a mature system is deployed, user simulators are able to provide task-oriented dialogues at a lower cost. There are two kinds of user simulators. One is agenda-based simulators (Schatzmann and Young 2009; Li et al. 2016d; Ultes et al. 2017), which only feed dialogue systems with the pre-defined user goal as a user message, without surface realization. Another is model-based simulators (Chandramohan et al. 2011; Asri et al. 2016), which generate user utterances using language models given constraint information.

5.2 Evaluation methods for open-domain dialogue systems

Evaluation of open-domain dialogue systems has long been a challenging problem. Unlike task-oriented systems, there is no clear metric like task completion rate or task completion cost. Both human and automatic evaluation methods are developed for ODD during these years. Human evaluation has been adopted by many works (Ritter et al. 2011; Shang et al. 2015; Sordoni et al. 2015b) to converse with and rate dialogue agents. However, human evaluation is not an ideal approach for that human labor is expensive and the evaluation results are highly subjective, varying from person to person. Researchers tend to hire crowdsource workers (Ritter et al. 2011; Shang et al. 2015; Sordoni et al. 2015b) or random people (Moon et al. 2019; Jung et al. 2020) to conduct human evaluation, both of which have two main drawbacks: 1. The evaluator group is highly random, and there exists a huge gap between people with different knowledge levels or from different domains. 2. Though individual bias could be weakened by increasing the number of evaluators, the evaluator group cannot be very large because of the limited budgets (in the articles mentioned above the sizes of human evaluator groups are usually 5–20). Thus, automatic and objective metrics are desirable. In general, there are two categories of automatic metrics in recent research: word-overlap metrics and neural metrics.

Word-overlap Metrics are widely used in Machine Translation and Summarization tasks, which calculate the similarity between the generated sequence and the ground truth sequence. Representative metrics like BLEU (Papineni et al. 2002) and ROUGE (Lin 2004) are n-gram matching metrics. METEOR (Banerjee and Lavie 2005) was further proposed with an improvement based on BLEU. It identified the paraphrases and synonyms between the generated sequence and the ground truth. Galley et al. (2015) extended the BLEU by exploiting numerical ratings of responses. Liu et al. (2016) argued that word-overlap metrics were not correlated well with human evaluation. These metrics are effective in Machine Translation because each source sentence has a ground truth to compare with, whereas in dialogues there may be many possible responses corresponding with one

user message, and thus an acceptable response may receive a low score if simply computing word-overlap metrics.

Neural Metrics are metrics computed by neural models. Neural methods improve the evaluation effectiveness in terms of adaptability compared with word-overlap metrics, but they require an additional training process. Su et al. (2015) used an RNN and a CNN model to extract turn-level features in a sequence and give the score. Tao et al. (2018) proposed Ruber, which was an automatic metric combining referenced and unreferenced components. The referenced one computed the similarity between generated response representations and ground truth representations, while the unreferenced one learned a scoring model to rate the query-response pairs. Lowe et al. (2017) learned representations of dialogue utterances using an RNN and then computed the dot-product between generated response and ground truth response as an evaluation score. Kannan and Vinyals (2017) and Bruni and Fernández (2017) used the discriminator of a GAN framework to distinguish the generated responses from human responses. If a generated response achieved a high confidence score, this was indicative of a human-like response, thus desirable.

Evaluation of open-domain dialogue systems is a hot topic at present and many researchers cast their eyes on this task recently. Some papers introduce two or more custom evaluation metrics for better evaluation, such as response diversity, response consistency, naturalness, knowledgeability, understandability, etc., to study "what to evaluate". Bao et al. (2019) evaluated the generated responses by designing two metrics. One was the informativeness metric calculating information utilization over turns. Another was the coherence metric, which was predicted by GRUs, given the response, context, and background as input. Likewise, Akama et al. (2020) designed scoring functions to compute connectivity of utterance pairs and content relatedness as two evaluation metrics and used another fusion function to combine the metrics. Pang et al. (2020) combined four metrics in their automatic evaluation framework: the context coherence metric based on GPT-2; phrase fluency metric based on GPT-2; diversity metric based on n-grams; logical self-consistency metric based on textual-entailment-inference. Mehri and Eskenazi (2020) proposed a reference-free evaluation metric. They annotated responses considering the following qualities: Understandable (0–1), Maintains Context (1–3), Natural (1–3), Uses Knowledge (0–1), Interesting (1–3), Overall Quality (1–5). Furthermore, a transformer was trained on these annotated dialogues to compute the score of quality.

Apart from "what to evaluate", there is also a multitude of papers studying "how to evaluate", which focus more on refining the evaluation process. Liang et al. (2020) proposed a three-stage framework to denoise the self-rating process. They first performed dialogue flow anomaly detection via self-supervised representation learning, and then the model was fine-tuned with smoothed self-reported user ratings. Finally, they performed a denoising procedure by calculating the Shapley value and removed the samples with negative values. Zhao et al. (2020a) trained RoBERTa as a response scorer to achieve reference-free and semi-supervised evaluation. Sato et al. (2020) constructed a test set by first generating several responses based on one user message and then human evaluation was performed to annotate each response with a score, where the response with the highest score was taken as a true response and the remainder taken as false responses. Dialogue systems were further evaluated by comparing the response selection accuracy on the test set, where a cross-entropy loss was calculated between the generated response and candidate responses to perform the selection operation. Likewise, Sinha et al. (2020) trained a BERT-based model to discriminate between true and false responses, where false responses were automatically generated. The model was further used to predict the evaluation score of a response based on dialogue context. Huang et al. (2020a) argued that responses should not be simply

evaluated based on their surface-level features, and instead the topic-level features were more essential. They incorporated a common-sense graph in their evaluation framework to obtain topic-level graph representations. The topic-level graph representation and utterance-level representation were jointly considered to evaluate the coherence of responses generated by open-domain dialogue systems.

The ranking is also an approach that evaluates dialogue systems effectively. Gao et al. (2020b) leveraged large-scale human feedback data such as upvotes, downvotes, and replies to learn a GPT-2-based response ranker. Thus, responses were evaluated by their rankings given by the ranker. Deriu et al. (2020) also evaluated the dialogue systems by ranking. They proposed a low-cost human-involved evaluation framework, in which different conversational agents conversed with each other and the human's responsibility was to annotate whether the generated utterance was human-like or not. The systems were evaluated by comparing the number of turns their responses were judged as human-like responses.

6 Datasets

The dataset is one of the most essential components in dialogue systems study. Nowadays the datasets are not enough no matter for task-oriented or open-domain dialogue systems, especially for those tasks requiring additional annotations (Novikova et al. 2017). For task-oriented dialogue systems, data can be collected via two main methods. One is to recruit human labor via crowdsourcing platforms to produce dialogues in a given task. Another is to collect dialogues in real task completions like film ticket booking. For open-domain dialogue systems, apart from dialogues collected in real interactions, social media is also a significant source of data. Some social media companies such as Twitter and Reddit provide API access to a small proportion of posts, but these services are restricted by many legal terms which affect the reproducibility of research. As a result, many recent works in dialogue systems collect their own datasets for train and test.

In this section, we review and categorize these datasets and make a comprehensive summary. To our best knowledge, Table 3 and 4 cover almost all available datasets used in recent task-oriented or open-domain dialogue systems.

6.1 Datasets for task-oriented dialogue systems

See Table 3.

Table 3 Datasets for Task-oriented dialogue systems

Name	Description	Task	Origin
Schema	A dataset mainly for dialogue state tracking	Dialogue State Tracking	Rastogi et al. (2020)
MetaLWOZ	Collected by crowdsourcing platforms, spanning over 227 tasks and 47 domains. This dataset is designed for learning in unseen domains	Domain Transfer	Lee et al. (2019)
E2E	A dataset for end-to-end dialogue generation in restaurant domain. Data is collected in crowdsourced fashion	End-to-end Task-oriented Dialogue Systems	Novikova et al. (2017)
MSR-E2E	Contain dialogues spanning over 3 domains: movie-ticket booking, restaurant reservation, and taxi booking	End-to-end Task-oriented Dialogue Systems	Li et al. (2018)
YELPNLG	A corpus consisting of utterances spanning over different restaurant attributes	Natural Language Generation	Oraby et al. (2019)
Clinical Conversation data set	It consists of conversations between physicians and participants	Natural Language Understanding	Du et al. (2019)
OOS	A large-scale dataset for intent detection.	Natural Language Understanding	Larson et al. (2019)
ATIS	A dataset consisting of voice calls from people who intend to make flight reservations	Natural Language Understanding; Dialogue State Tracking	Tur et al. (2010)
MultiWOZ	Human–human written conversations with rich annotations spanning over multi-domains	Task-oriented Dialogue	Budzianowski et al. (2018)
SNIPS-NLU	Task-oriented dialogue dataset colleted in a crowd-sourced fashion. It was used to train voice assistant agents.	Task-oriented Dialogue	https://github.com/snipsco/nlubenchmark
bAbI	Restaurant table reservation dialogues	Task-oriented Dialogue	Bordes et al. (2017)
JDC	A Chinese customer service dataset, consisting of context-response pairs	Task-oriented Dialogue	https://www.jddc.jd.com
UbuntuV2	It consists of dialogues collected via Ubuntu question-answering forum	Task-oriented Dialogue	Lowe et al. (2015)
MICROSOFT DIALOGUE CHALLENGE data set	A task-oriented dataset collected via Amazon Mechanical Turk	Task-oriented Dialogue	Li et al. (2018)

Table 3 (continued)

Name	Description	Task	Origin
WOZ	Task-oriented data collected in crowdsourced fashion.	Task-oriented Dialogue	Wen et al. (2017)
DSTC series	Multi-domain task-oriented dataset	Task-oriented Dialogue	https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge/
SimDial	Simulated conversations spanning over multiple domains.	Task-oriented Dialogue	Zhao and Eskenazi (2018)
SMD	Human-human dialogues in weather, navigation and scheduling domain	Task-oriented Dialogue	Eric et al. (2017)
BANKING	Question-answer pairs with 77 categories in e-banking domain	Task-oriented Dialogue	Henderson et al. (2019b)
Weather forecast	A task-oriented dataset in the weather domain	Task-oriented Dialogue	Balakrishnan et al. (2019)
MedDialog-(EN,CN)	Large scale dataset in medical domain consisting of conversations between doctors and patients	Task-oriented Dialogue	He et al. (2020b)
CamRest	It consists of human-human multi-turn dialogues in restaurant domain	Task-oriented Dialogue	Wen et al. (2016a)
Taskmaster	Contain dialogues spanning over 6 domains. It has 22.9 average length of conversational turns	Task-oriented Dialogue	Byrne et al. (2019)
Frames	Conversational dataset with annotations of semantic frame tracking	Task-oriented Dialogue	El Asri et al. (2017)
JDDC	A Chinese customer service dataset, consisting of context-response pairs	Task-oriented Dialogue	Chen et al. (2020b)
Court Debate Dataset	A task-oriented dataset in judicial field containing court debate conversations	Task-oriented Dialogue	Ji et al. (2020)
TreeDST	A task-oriented dataset annotated with tree structured dialogue states and agent acts	Task-oriented Dialogue	Cheng et al. (2020)
RiSAWOZ	Contain utterances for 12 domains, annotated with rich semantic information	Task-oriented Dialogue	Quan et al. (2020)

Table 3 (continued)

Name	Description	Task	Origin
Cambridge Restaurant	A task-oriented dataset in restaurant booking field	Task-oriented Dialogue	Wen et al. (2017)
SB-TOP	A task-oriented dataset with semantic parsing annotation. It spans over 4 domains: Reminder, Weather, Calling and Music	Task-oriented Dialogue	Aghajanyan et al. (2020)
GSIM	A machine-machine task-oriented dataset. It covers two domains: restaurant table booking and movie ticket booking	Task-oriented Dialogue	Shah et al. (2018)
SGD	A schema-guided dataset spanning over multiple domains	Task-oriented Dialogue	Rastogi et al. (2020)
cite-8K	A task-oriented dataset collected in restaurant booking calls.	Task-oriented Dialogue	Coope et al. (2020)

6.2 Datasets for open-domain dialogue systems

See Table 4.

Table 4 Datasets for open-domain dialogue systems

Name	Description	Task	Origin
Large-Scale Corpus for Conversation Disentanglement	A dataset consisting of messages annotated with reply-structure graphs for dialogue disentanglement	Conversation Disentaglement	Kummerfeld et al. (2019)
DuConv	Collected in conversations between a conversation leader and a conversation follower	Conversation Topic	Wu et al. (2019c)
PERSUASION FOR GOOD	A topic-oriented dataset annotated with persuasion strategies	Conversation Topic	Wang et al. (2019b)
MutualFriends	A topic-oriented dataset based on bot-bot stratigical conversations.	Conversation Topic	He et al. (2017)
SAMSum	A large-scale dialogue summary dataset	Conversation Topic	Gliwa et al. (2019)
OpenDialKG	It consists conversations between two agents and each dialogue corresponds with a knowledge graph path annotation	Conversation Topic; Dialogue Reasoning	Moon et al. (2019)
doc2dial	A dataset consisting of conversations annotated with goals and accociated documents	Conversation Topic; Knowledge-Grounded System	Feng et al. (2020c)
DialEdit	A dataset constructed for image editing via conversational language instructions.	Conversational Image Editing	Manuvirakurike et al. (2018)
CHART DIALOGS	A dataset containing dialogues describing matplotlib plot features.	Conversational Plotting	Shao and Nakashole (2020)
CONAN	A multilingual dataset for hate speech tackling.	Dialogue Classification	Chung et al. (2019)
Dialogue NLI	A NLI dataset with sentences annotated with entailment (E), neutral (N), or contradiction (C)	Dialogue Inference	Welleck et al. (2019)

Table 4 (continued)

Name	Description	Task	Origin
MuTual	A dialogue reasoning dataset containing English listening comprehension exams	Dialogue Reasoning	Cui et al. (2020)
RST-DT	It consists of samples from 385 news articles annotated with dialogue features.	Discourse Parsing	Carlson et al. (2002)
NLPCC	A dataset consisting of emotional classification data	Empathetic Response	http://tcci.ccf.org.cn/nlpcc.php
MELD	A multi-party conversational dataset with emotion annotations.	Empathetic Response	Poria et al. (2019)
EMPATHETIC DIALOGUES	A dataset containing conversations annotated with emotion labels	Empathetic Response	Rashkin et al. (2019)
IEMOCAP	Contain multi-party dialogues. Each dialogue is annotated with an emotion label	Empathetic Response	Busso et al. (2008)
EmoryNLP	Collected from Friends' TV series, annotated with emotion labels	Empathetic Response	Zahiri and Choi (2017)
MojiTalk	A largescale dataset collected from Twitter, including emojis	Empathetic Response	Zhou and Wang (2018)
CBET	A dialogue dataset annotated with nine emotion labels: surprise, anger, love, sadness, joy, fear, guilt, disgust, and thankfulness	Empathetic Response	Yadollahi et al. (2017)
Stanford Politeness Corpus	A conversational dataset annotated with politeness labels.	Empathetic Response	Danescu-Niculescu-Mizil et al. (2013)
AIT-2018	Collected in SemEval-2018 Task 1: Affect in Tweets	Empathetic Response	Mohammad et al. (2018)
EMOTyDA	A dataset containing short videos about multi-party conversations, each annotated with respective emotions.	Empathetic Response; Visual Dialogue	Saha et al. (2020)
Wizard of Wikipedia	A large-scale dataset consisting of conversations grounded with Wikipedia knowledge	Knowledge-Grounded System	Dinan et al. (2019b)

Table 4 (continued)

Name	Description	Task	Origin
CMU DoG	A dataset consists of conversations grounded with Wikipedia articles about popular movies.	Knowledge-Grounded System	Zhou et al. (2018)
Holl-E	Contain dialogues grounded with documents.	Knowledge-Grounded System	Moghe et al. (2018)
Interview	A dataset containing multi-party conversations in the form of interviews	Knowledge-Grounded System	Majumder et al. (2020b)
Curiosity	An open-domain dataset annotated with pre-existing user knowledge and dialogue acts, also grounding in Wikipedia	Knowledge-Grounded System	Rodriguez et al. (2020)
KdConv	A chinese knowledge-grounded dialogue dataset.	Knowledge-Grounded System	Zhou et al. (2020a)
ELI5	A QA dataset grounded with retrieved documents.	Knowledge-Grounded System	Fan et al. (2019)
Topical Chat	A knowledge-grounded dataset where the knowledge spans over eight different topics.	Knowledge-Grounded System; Conversation Topic	Gopalakrishnan et al. (2019)
WHERE ARE YOU?	A dialogue dataset annotated with localization information	Localization Dialogue	Hahn et al. (2020)
MMD	A multi-modal dataset consisting of dialogues between sales agents and shoppers	Multi-modal Dialogue	Saha et al. (2018)
OpenSubtitles	A multilingual dataset made up of movie captions, containing about 8 billion words	Open-domain Dialogue	Tiedemann (2012)
NTCIR	A social media dataset collected from Sina Weibo.	Open-domain Dialogue	http://research.nii.ac.jp/ntcir/data/data-en.html
Twitter	A social media dataset collected from Twitter	Open-domain Dialogue	https://github.com/Marsan-Ma-zz/chat corpus
Douban Conversation Corpus	A social media dataset collected from Douban	Open-domain Dialogue	Zhang et al. (2018c)

Table 4 (continued)

Name	Description	Task	Origin
E-commerce Dialogue Corpus	It consists of conversations between customers and customer service staff on Taobao.	Open-domain Dialogue	Zhang et al. (2018c)
REDDIT	A social media dataset collected from REDDIT	Open-domain Dialogue	Henderson et al. (2019a)
STC-SeFun	A social media dataset collected from Tieba, Zhidao, Douban and Weibo	Open-domain Dialogue	Bi et al. (2019)
DailyDialog	A dataset consisting of daily dialogues, annotated with conversation intention and emotion information	Open-domain Dialogue	Li et al. (2017e)
PDTB	Dialogue dataset annotated with discourse relations	Open-domain Dialogue	Miltsakaki et al. (2004)
Luna	Dialogue dataset with Italian relation annotations.	Open-domain Dialogue	Tonelli et al. (2010)
Edina-DR	Dialogue dataset with English relation annotations, which is based on Luna data set	Open-domain Dialogue	Ma et al. (2019)
Cornell Movie Dialog Corpus	A dialogue dataset collected via IMDB database	Open-domain Dialogue	Danescu-Niculescu-Mizil and Lee (2011)
Reddit Movie Dialogue Dataset	A movie dialogue dataset collected from Reddit	Open-domain Dialogue	Liu et al. (2020a)
LIGHT	A dialogue dataset with configurable text adventure environment	Open-domain Dialogue	Urbanek et al. (2019)
This American Life	A media dialogue dataset collected in long-form expository podcast episodes	Open-domain Dialogue	Mao et al. (2020)
RadioTalk	A media dialogue dataset collected from radio transcripts	Open-domain Dialogue	Beeferman et al. (2019)
French EPAC	A media dialogue dataset collected from news	Open-domain Dialogue	Estève et al. (2010)
TREC Conversational Assistance	An open-domain dataset spanning 30 conversation topics.	Open-domain Dialogue	Dalton et al. (2020)

Table 4 (continued)

Name	Description	Task	Origin
Search as a Conversation	A dataset for conversations with search engines.	Open-domain Dialogue	Ren et al. (2020)
Amazon Alexa Prize Competition	A dataset containing real-world conversations between Amazon Alexa customers and Gunrock, which is a champion chatbot	Open-domain Dialogue	Ram et al. (2018)
SwitchBoard	An open-domain dataset containing English phone conversations	Open-domain Dialogue	Jurafsky (1997)
Zhihu	A Chinese social media dataset with posts and comments	Open-domain Dialogue	https://www.zhihu.com
SPOLIN	A dataset containing yes-and conversations	Open-domain Dialogue	Cho and May (2020)
CRD3	A dataset collected in the role-playing game Dungeons and Dragons	Open-domain Dialogue	Rameshkumar and Bailey (2020)
Baidu Zhidao	A Chinese social media dataset with posts and comments	Open-domain Dialogue	https://zhidao.baidu.com/
Webis Gmane Email Corpus 2019	A conversational dataset collected from 153M emails	Open-domain Dialogue	Bevendorff et al. (2020)
LibreSpeech Corpus	Contain 500 hours' speech produced by 1252 participants	Open-domain Dialogue	Panayotov et al. (2015)
Motivational Interviewing	A dialogue dataset about conversational psychotherapy	Open-domain Dialogue	Tanana et al. (2016)
SubTle Corpus	Contact Ameixa for data	Open-domain Dialogue	Lubis et al. (2018)
TED-LIUM	TED-talk monologues	Open-domain Dialogue	Fung et al. (2016)
ECG NLPCC 2017 Data	Conversational dataset extracted from Weibo	Open-domain Dialogue	Huang et al. (2018)
SEMEVAL15	QA dataset with answer quality annotations via Amazon Mechanical Turk	Question Answering	Nakov et al. (2015)
AMAZONQA	A QA dataset solving one-to-many problems	Question Answering	Wan and McAuley (2016)

Table 4 (continued)

Name	Description	Task	Origin
TGIF-QA	A video-grounded QA dataset	Question Answering	Jang et al. (2017)
QuAC	A QA dataset with 14K QA dialogues	Question Answering	Choi et al. (2018)
SQuAD	A question-answering dataset collected in crowdsourced fashion	Question Answering	Rajpurkar et al. (2018)
LIF	A dataset constructed based on QuAC	Question Answering	Kundu et al. (2020)
Yelp	It consists of customer reviews from Yelp Dataset Challenge	Response Retrieval	Tang et al. (2015)
Debates	The dataset consists of debates on Congressional bills.	Response Retrieval	Thomas et al. (2006)
PERSONACHAT	It provides profile information of the agents and background of users.	Speaker Consistency and Personality Response	Zhang et al. (2018a)
KvPI	Contain consistency annotations between response and corresponding key-value profiles	Speaker Consistency and Personality Response	Song et al. (2020b)
ConvAI2	A dataset constructed on the base of Persona-Chat, each conversation having profiles from a set containing persona candidates.	Speaker Consistency and Personality Response	Dinan et al. (2019a)
PEC	An open-domain dataset annotated with persona labels.	Speaker Consistency and Personality Response; Empathetic Response	Zhong et al. (2020)
GuessWhat?!	A visual dialogue dataset for a two-player game about object recognition.	Visual Dialogue	de Vries et al. (2017)
VisDial	A visual dialogue dataset whose images are obtained from COCO data set.	Visual Dialogue	https://visualdialog.org/ data
AVSD	A video-grounded dialogue dataset.	Visual Dialogue	Yoshino et al. (2018)
VFD	A visual dialogue dataset annotated with unique eye-gaze locations.	Visual Dialogue	Kamezawa et al. (2020)
PhotoBook	A dataset for task-oriented visual dialogues.	Visual Dialogue	Haber et al. (2019)

Table 4 (continued)

Name	Description	Task	Origin
IGC	A dataset containing conversations discussing a given image	Visual Dialogue	Mostafazadeh et al. (2017)
Image-Chat	Contain conversations grounded with images. The conversations are also annotated with personality	Visual Dialogue; Speaker Consistency and Personality Response	Shuster et al. (2020b)

7 Conclusions and trends

More and more researchers are investigating conversational tasks. One factor contributing to the popularity of conversational tasks is the increasing demand for chatbots in industry and daily life. Industry agents like Apple's Siri, Microsoft's Cortana, Facebook M, Google Assistant, and Amazon's Alexa have brought huge convenience to people's lives. Another reason is that a considerable amount of natural language data is in the form of dialogues, which contributes to the efforts in dialogue research.

In this paper we discuss dialogue systems from two perspectives: model and system type. Dialogue systems are a complicated but promising task because it involves the whole process of communication between agent and human. The works of recent years show an overwhelming preference for neural methods, no matter in task-oriented or open-domain dialogue systems. Neural methods outperform traditional rule-based methods, statistical methods, and machine learning methods for that neural models have stronger fitting ability and require less hand-crafted feature engineering.

We systematically summarized and categorized the latest works in dialogue systems, and also in other dialogue-related tasks. We hope these discussions and insights provide a comprehensive picture of the state-of-the-art in this area and pave the way for further research. Finally, we discuss some possible research trends arising from the works reviewed:

Multimodal dialogue systems The world is multimodal and humans observe it via multiple senses such as vision, hearing, smell, taste, and touch. In a conversational interaction, humans tend to make responses not only based on text, but also on what they see and hear. Thus, some researchers argue that chatbots should also have such ability to blend information from different modalities. There are some recent works trying to build multimodal dialogue systems (Le et al. 2019; Chauhan et al. 2019; Saha et al. 2020; Singla et al. 2020; Young et al. 2020), but these systems are still far from mature.

Multitask dialogue systems Dialogue systems are categorized into task-oriented and open-domain systems. Such a research boundary has existed for a long time because task-oriented dialogue systems involve dialogue states, which constrain the decoding process. However, works in end-to-end task-oriented dialogue systems and knowledge-grounded open-domain systems provide a possibility of blending these two categories into a single framework, or even a single model. Such blended dialogue systems perform as assistants and chatbots simultaneously (Young et al. 2022).

Corpus exploration on Internet In Sect. 6 we reviewed many datasets for dialogue systems training. However, data is still far from enough to train a perfect dialogue system. Many learning techniques are designed to alleviate this problem, such as reinforcement learning, meta-learning, transfer learning, and active learning. But many works ignore a significant source of information, which is the dialogue corpus on the Internet. There is a large volume of conversational corpus on the Internet but people have no access to the raw corpus because much of it is in a messy condition. In the future, dialogue agents should be able to explore useful corpus on the Internet in real-time for training. This can be achieved by standardizing online corpus access and their related legal terms. Moreover, real-time conversational corpus exploration can be an independent task that deserves study.

User modeling User modeling is a hot topic in both dialogue generation (Gür et al. 2018; Serras et al. 2019) and dialogue systems evaluation (Kannan and Vinyals 2017).

Basically, the user modeling module tries to simulate the real decisions and actions of a human user. It makes decisions based on the dialogue state or dialogue history. In dialogue generation tasks, modeling the user helps the agent converse more coherently, based on the background information or even speaking habits. Besides that, a mature user simulator can provide an interactive training environment, which reduces the reliance on annotated training samples when training a dialogue system. In dialogue systems evaluation tasks, a user simulator provides user messages to test a dialogue agent. More recent user simulators also give feedback concerning the responses generated by the dialogue agent. However, user modeling is a challenging task since no matter whether explicit user simulation or implicit user modeling is actually the same in difficulty as response generation. Since response generation systems are not perfect yet, user modeling can still be a topic worthy of study.

Dialogue generation with a long-term goal Most of our daily conversations are chitchats without any purpose. However, there are quite a few scenarios when we purposely guide the conversation content to achieve a specific goal. Current open-domain dialogue systems tend to model the conversation without a long-term goal, which does not exhibit enough intelligence. There are some recent works that apply reinforcement policy learning to model a long-term reward that encourages the agent to converse with a long-term goal, such as the work of Xu et al. (2020a). This topic will lead to strong artificial intelligence, which is useful in some real-life applications such as negotiation or story-telling chatbots.

Acknowledgements This research/project is supported by A*STAR under its Industry Alignment Fund (LOA Award I1901E0046).

References

Abro WA, Qi G, Ali Z, Feng Y, Aamir M (2020) Multi-turn intent determination and slot filling with neural networks and regular expressions. Knowl-Based Syst 208:106428

Abro WA, Aicher A, Rach N, Ultes S, Minker W, Qi G (2022) Natural language understanding for argumentative dialogue systems in the opinion building domain. Knowl-Based Syst 242:108318

Agarwal S, Bui T, Lee JY, Konstas I, Rieser V (2020) History for visual dialog: Do we really need it? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, pp 8182–8197, https://doi.org/10.18653/v1/2020.acl-main.728

Aghajanyan A, Maillard J, Shrivastava A, Diedrick K, Haeger M, Li H, Mehdad Y, Stoyanov V, Kumar A, Lewis M, Gupta S (2020) Conversational semantic parsing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 5026–5035, https://doi.org/10.18653/v1/2020.emnlp-main.408

Akama R, Yokoi S, Suzuki J, Inui K (2020) Filtering noisy dialogue corpora by connectivity and content relatedness. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 941–958, https://doi.org/10.18653/v1/2020.emnlp-main.68

Alberti C, Ling J, Collins M, Reitter D (2019) Fusion of detected objects in text for visual question answering. 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), association for computational linguistics, Hong Kong, China, pp 2131–2140, https://doi.org/10.18653/v1/D19-1219

Aloysius N, Geetha M (2017) A review on deep convolutional neural networks. In: 2017 international conference on communication and signal processing (ICCSP), IEEE, pp 0588–0592

Arora S, Batra K, Singh S (2013) Dialogue system: a brief review. arXiv:1306.4134

Asghar N, Poupart P, Jiang X, Li H (2017) Deep active learning for dialogue generation. In: Proceedings of the 6th joint conference on lexical and computational semantics (SEM 2017), association for computational linguistics, Vancouver, Canada, pp 78–83, https://doi.org/10.18653/v1/S17-1008

Asri LE, He J, Suleman K (2016) A sequence-to-sequence model for user simulation in spoken dialogue systems. In: Morgan N (ed) Interspeech 2016, 17th annual conference of the international speech communication association, San Francisco, CA, USA, September 8–12, 2016, ISCA, pp 1151–1155, https://doi.org/10.21437/Interspeech.2016-1175

Aubert X, Dugast C, Ney H, Steinbiss V (1994) Large vocabulary continuous speech recognition of wall street journal data. In: Proceedings of ICASSP'94. IEEE International conference on acoustics, speech and signal processing, IEEE, vol 2, pp II–129

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, arXiv:1409.0473

Baheti A, Ritter A, Small K (2020) Fluent response generation for conversational question answering. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 191–207, https://doi.org/10.18653/v1/2020.acl-main.19

Balakrishnan A, Rao J, Upasani K, White M, Subba R (2019) Constrained decoding for neural NLG from compositional representations in task-oriented dialogue. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 831–844, https://doi.org/10.18653/v1/P19-1080

Banerjee S, Lavie A (2005) METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, association for computational linguistics, Ann Arbor, Michigan, pp 65–72, https://aclanthology.org/W05-0909

Bao S, He H, Wang F, Lian R, Wu H (2019) Know more about each other: Evolving dialogue strategy via compound assessment. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5382–5391, https://doi.org/10.18653/v1/P19-1535

Bao S, He H, Wang F, Wu H, Wang H (2020) PLATO: Pre-trained dialogue generation model with discrete latent variable. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 85–96, https://doi.org/10.18653/v1/2020.acl-main.9

Bapna A, Tür G, Hakkani-Tür D, Heck LP (2017) Towards zero-shot frame semantic parsing for domain scaling. In: Lacerda F (ed) Interspeech 2017, 18th annual conference of the international speech communication association, Stockholm, Sweden, August 20-24, 2017, ISCA, pp 2476–2480, http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0518.html

Beeferman D, Brannon W, Roy D (2019) Radiotalk: A large-scale corpus of talk radio transcripts. In: Kubin G, Kacic Z (eds) Interspeech 2019, 20th annual conference of the international speech communication association, Graz, Austria, 15–19 September 2019, ISCA, pp 564–568, https://doi.org/10.21437/Interspeech.2019-2714

Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166

Bevendorff J, Al Khatib K, Potthast M, Stein B (2020) Crawling and preprocessing mailing lists at scale for dialog analysis. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1151–1158, https://doi.org/10.18653/v1/2020.acl-main.108

Bi W, Gao J, Liu X, Shi S (2019) Fine-grained sentence functions for short-text conversation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational Linguistics, Florence, Italy, pp 3984–3993, https://doi.org/10.18653/v1/P19-1389

Bordes A, Usunier N, García-Durán A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems 26: 27th annual conference on neural information processing systems 2013. Proceedings of a meeting held December 5–8, 2013, Lake Tahoe, Nevada, United States, pp 2787–2795, https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

Bordes A, Glorot X, Weston J, Bengio Y (2014) A semantic matching energy function for learning with multi-relational data. Mach Learn 94(2):233–259

Bordes A, Boureau Y, Weston J (2017) Learning end-to-end goal-oriented dialog. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, https://openreview.net/forum?id=S1Bb3D5gg

Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y (2019) COMET: Commonsense transformers for automatic knowledge graph construction. In: Proceedings of the 57th Annual

Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 4762–4779, https://doi.org/10.18653/v1/P19-1470
Bouchacourt D, Baroni M (2019) Miss tools and mr fruit: Emergent communication in agents learning about object affordances. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3909–3918, https://doi.org/10.18653/v1/P19-1380
Boyd A, Puri R, Shoeybi M, Patwary M, Catanzaro B (2020) Large scale multi-actor generative dialog modeling. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, Online, pp 66–84, https://doi.org/10.18653/v1/2020.acl-main.8
Bruni E, Fernández R (2017) Adversarial evaluation for open-domain dialogue generation. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics, Saarbrücken, Germany, pp 284–288, https://doi.org/10.18653/v1/W17-5534
Budzianowski P, Wen TH, Tseng BH, Casanueva I, Ultes S, Ramadan O, Gašić M (2018) MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics, Brussels, Belgium, pp 5016–5026, https://doi.org/10.18653/v1/D18-1547
Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) Iemocap: Interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359
Byrne B, Krishnamoorthi K, Sankar C, Neelakantan A, Goodrich B, Duckworth D, Yavuz S, Dubey A, Kim KY, Cedilnik A (2019) Taskmaster-1: Toward a realistic and diverse dialog dataset. 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), association for computational linguistics, Hong Kong, China, pp 4516–4525, https://doi.org/10.18653/v1/D19-1459
Cahill L, Doran C, Evans R, Mellish C, Paiva D, Reape M, Scott D, Tipper N (1999) In search of a reference architecture for nlg systems. In: Proceedings of the 7th European workshop on natural language generation, Citeseer, pp 77–85
Campagna G, Foryciarz A, Moradshahi M, Lam M (2020) Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 122–132, https://doi.org/10.18653/v1/2020.acl-main.12
Cao J, Tanana M, Imel Z, Poitras E, Atkins D, Srikumar V (2019) Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5599–5611, https://doi.org/10.18653/v1/P19-1563
Carlson L, Okurowski ME, Marcu D (2002) RST discourse treebank. Linguistic Data Consortium, University of Pennsylvania
Casanueva I, Temčinas T, Gerz D, Henderson M, Vulić I (2020) Efficient intent detection with dual sentence encoders. In: Proceedings of the 2nd workshop on natural language processing for conversational AI, association for computational linguistics, online, pp 38–45, https://doi.org/10.18653/v1/2020.nlp4convai-1.5
Chandramohan S, Geist M, Lefevre F, Pietquin O (2011) User simulation in dialogue systems using inverse reinforcement learning. In: Twelfth annual conference of the international speech communication association
Chauhan H, Firdaus M, Ekbal A, Bhattacharyya P (2019) Ordinal and attribute aware response generation in a multimodal dialogue system. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5437–5447, https://doi.org/10.18653/v1/P19-1540
Chen H, Liu X, Yin D, Tang J (2017) A survey on dialogue systems: Recent advances and new frontiers. Acm Sigkdd Explorations Newslett 19(2):25–35
Chen J, Yang D (2020) Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 4106–4118, https://doi.org/10.18653/v1/2020.emnlp-main.336
Chen J, Zhang R, Mao Y, Xu J (2020a) Parallel interactive networks for multi-domain dialogue state generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 1921–1931, https://doi.org/10.18653/v1/2020.emnlp-main.151

Chen L, Zhou X, Chang C, Yang R, Yu K (2017b) Agent-aware dropout DQN for safe and efficient online dialogue policy learning. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, Copenhagen, Denmark, pp 2454–2464, https://doi.org/10.18653/v1/D17-1260

Chen M, Liu R, Shen L, Yuan S, Zhou J, Wu Y, He X, Zhou B (2020b) The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service. In: Proceedings of the 12th language resources and evaluation conference, European language resources association, Marseille, France, pp 459–466, https://aclanthology.org/2020.lrec-1.58

Chen W, Chen J, Qin P, Yan X, Wang WY (2019a) Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3696–3709, https://doi.org/10.18653/v1/P19-1360

Chen X, Xu J, Xu B (2019b) A working memory model for task-oriented dialog response generation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 2687–2693, https://doi.org/10.18653/v1/P19-1258

Chen X, Meng F, Li P, Chen F, Xu S, Xu B, Zhou J (2020c) 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), association for computational linguistics, online, pp 3426–3437, https://doi.org/10.18653/v1/2020.emnlp-main.275

Chen Y, Hakkani-Tür D, Tür G, Gao J, Deng L (2016) End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In: Morgan N (ed) Interspeech 2016, 17th Annual conference of the international speech communication association, San Francisco, CA, USA, September 8–12, 2016, ISCA, pp 3245–3249, https://doi.org/10.21437/Interspeech.2016-312

Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu J (2019c) Uniter: Learning universal image-text representations. ECCV

Cheng J, Agrawal D, Martínez Alonso H, Bhargava S, Driesen J, Flego F, Kaplan D, Kartsaklis D, Li L, Piraviperumal D, Williams JD, Yu H, Ó Séaghdha D, Johannsen A (2020) Conversational semantic parsing for dialog state tracking. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 8107–8117, https://doi.org/10.18653/v1/2020.emnlp-main.651

Cho H, May J (2020) Grounding conversations with improvised dialogues. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2398–2413, https://doi.org/10.18653/v1/2020.acl-main.218

Cho K, van Merriënboer B, Bahdanau D, Bengio Y (2014a) On the properties of neural machine translation: Encoder–decoder approaches. In: Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation, association for computational linguistics, Doha, Qatar, pp 103–111, https://doi.org/10.3115/v1/W14-4012

Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014b) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, Doha, Qatar, pp 1724–1734, https://doi.org/10.3115/v1/D14-1179

Choi E, He H, Iyyer M, Yatskar M, Yih Wt, Choi Y, Liang P, Zettlemoyer L (2018) QuAC: Question answering in context. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics, Brussels, Belgium, pp 2174–2184, https://doi.org/10.18653/v1/D18-1241

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555

Chung YL, Kuzmenko E, Tekiroglu SS, Guerini M (2019) CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 2819–2829, https://doi.org/10.18653/v1/P19-1271

Cogswell M, Lu J, Jain R, Lee S, Parikh D, Batra D (2020) Dialog without dialog data: Learning visual dialog agents from VQA data. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, https://proceedings.neurips.cc/paper/2020/hash/e7023ba77a45f7e84c5ee8a28dd63585-Abstract.html

Conneau A, Schwenk H, Barrault L, Lecun Y (2017) Very deep convolutional networks for text classification. In: Proceedings of the 15th conference of the European chapter of the association for

computational linguistics: volume 1, long papers, association for computational linguistics, Valencia, Spain, pp 1107–1116, https://aclanthology.org/E17-1104

Coope S, Farghly T, Gerz D, Vulić I, Henderson M (2020) Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, Online, pp 107–121, https://doi.org/10.18653/v1/2020.acl-main.11

Csáky R, Purgai P, Recski G (2019) Improving neural conversational models with entropy-based data filtering. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5650–5669, https://doi.org/10.18653/v1/P19-1567

Cui L, Wu Y, Liu S, Zhang Y, Zhou M (2020) MuTual: A dataset for multi-turn dialogue reasoning. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1406–1416, https://doi.org/10.18653/v1/2020.acl-main.130

Dai Y, Li H, Tang C, Li Y, Sun J, Zhu X (2020) Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 609–618, https://doi.org/10.18653/v1/2020.acl-main.57

Dai Z, Yang Z, Yang Y, Carbonell J, Le Q, Salakhutdinov R (2019) Transformer-XL: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 2978–2988, https://doi.org/10.18653/v1/P19-1285

Dalton J, Xiong C, Callan J (2020) Trec cast 2019: The conversational assistance track overview. http://arxiv.org/abs/2003.13624

Danescu-Niculescu-Mizil C, Lee L (2011) Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In: Proceedings of the 2nd workshop on cognitive modeling and computational linguistics, association for computational linguistics, Portland, Oregon, USA, pp 76–87, https://aclanthology.org/W11-0609

Danescu-Niculescu-Mizil C, Sudhof M, Jurafsky D, Leskovec J, Potts C (2013) A computational approach to politeness with application to social factors. In: Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Sofia, Bulgaria, pp 250–259, https://aclanthology.org/P13-1025

Deng L, Tur G, He X, Hakkani-Tur D (2012) Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In: 2012 IEEE spoken language technology workshop (SLT), IEEE, pp 210–215

Deoras A, Sarikaya R (2013) Deep belief network based semantic taggers for spoken language understanding. In: Interspeech, pp 2713–2717

Deriu J, Tuggener D, von Däniken P, Campos JA, Rodrigo A, Belkacem T, Soroa A, Agirre E, Cieliebak M (2020) Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3971–3984, https://doi.org/10.18653/v1/2020.emnlp-main.326

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), association for computational linguistics, Minneapolis, Minnesota, pp 4171–4186, https://doi.org/10.18653/v1/N19-1423

Dhingra B, Li L, Li X, Gao J, Chen YN, Ahmed F, Deng L (2017) Towards end-to-end reinforcement learning of dialogue agents for information access. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Vancouver, Canada, pp 484–495, https://doi.org/10.18653/v1/P17-1045

Dinan E, Logacheva V, Malykh V, Miller A, Shuster K, Urbanek J, Kiela D, Szlam A, Serban I, Lowe R, et al. (2019a) The second conversational intelligence challenge (convai2). https://arxiv.org/abs/1902.00098

Dinan E, Roller S, Shuster K, Fan A, Auli M, Weston J (2019b) Wizard of wikipedia: Knowledge-powered conversational agents. In: 7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, https://openreview.net/forum?id=r1l73iRqKm

Dong L, Huang S, Wei F, Lapata M, Zhou M, Xu K (2017) Learning to generate product reviews from attributes. In: Proceedings of the 15th conference of the European chapter of the association for

computational linguistics: volume 1, long papers, association for computational linguistics, Valencia, Spain, pp 623–632, https://aclanthology.org/E17-1059

Du N, Chen K, Kannan A, Tran L, Chen Y, Shafran I (2019) Extracting symptoms and their status from clinical conversations. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 915–925, https://doi.org/10.18653/v1/P19-1087

Du W, Black AW (2019) Boosting dialog response generation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 38–43, https://doi.org/10.18653/v1/P19-1005

Dušek O, Jurčíček F (2016a) A context-aware natural language generator for dialogue systems. In: Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue, association for computational linguistics, Los Angeles, pp 185–190, https://doi.org/10.18653/v1/W16-3622

Dušek O, Jurčíček F (2016b) Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers), association for computational linguistics, Berlin, Germany, pp 45–51, https://doi.org/10.18653/v1/P16-2008

El Asri L, Schulz H, Sharma S, Zumer J, Harris J, Fine E, Mehrotra R, Suleman K (2017) Frames: a corpus for adding memory to goal-oriented dialogue systems. In: Proceedings of the 18th annual sigdial meeting on discourse and dialogue, association for computational linguistics, Saarbrücken, Germany, pp 207–219, https://doi.org/10.18653/v1/W17-5526

Elder H, O'Connor A, Foster J (2020) How to make neural natural language generation as reliable as templates in task-oriented dialogue. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 2877–2888, https://doi.org/10.18653/v1/2020.emnlp-main.230

Elman JL (1990) Finding structure in time. Cogn Sci 14(2):179–211

Eric M, Krishnan L, Charette F, Manning CD (2017) Key-value retrieval networks for task-oriented dialogue. In: Proceedings of the 18th annual SIGdial meeting on discourse and dialogue, association for computational linguistics, Saarbrücken, Germany, pp 37–49, https://doi.org/10.18653/v1/W17-5506

Estève Y, Bazillon T, Antoine JY, Béchet F, Farinas J (2010) The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news. In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, http://www.lrec-conf.org/proceedings/lrec2010/pdf/650_Paper.pdf

Fan A, Jernite Y, Perez E, Grangier D, Weston J, Auli M (2019) ELI5: Long form question answering. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3558–3567, https://doi.org/10.18653/v1/P19-1346

Fan M, Zhou Q, Chang E, Zheng TF (2014) Transition-based knowledge graph embedding with relational mapping properties. In: Proceedings of the 28th Pacific asia conference on language, information and computing, department of linguistics, Chulalongkorn University, Phuket, Thailand, pp 328–337, https://aclanthology.org/Y14-1039

Feldman Y, El-Yaniv R (2019) Multi-hop paragraph retrieval for open-domain question answering. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 2296–2309, https://doi.org/10.18653/v1/P19-1222

Feng J, Tao C, Wu W, Feng Y, Zhao D, Yan R (2019) Learning a matching model with co-teaching for multi-turn response selection in retrieval-based dialogue systems. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3805–3815, https://doi.org/10.18653/v1/P19-1370

Feng S, Chen H, Li K, Yin D (2020a) Posterior-gan: Towards informative and coherent response generation with posterior generative adversarial network. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, pp 7708–7715, https://aaai.org/ojs/index.php/AAAI/article/view/6273

Feng S, Ren X, Chen H, Sun B, Li K, Sun X (2020b) Regularizing dialogue generation by imitating implicit scenarios. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 6592–6604, https://doi.org/10.18653/v1/2020.emnlp-main.534

Feng S, Wan H, Gunasekara C, Patel S, Joshi S, Lastras L (2020c) doc2dial: A goal-oriented document-grounded dialogue dataset. In: Proceedings of the 2020 conference on empirical methods in natural

language processing (EMNLP), association for computational linguistics, online, pp 8118–8128, https://doi.org/10.18653/v1/2020.emnlp-main.652
Ferracane E, Durrett G, Li JJ, Erk K (2019) Evaluating discourse in structured text representations. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 646–653, https://doi.org/10.18653/v1/P19-1062
Ficler J, Goldberg Y (2017) Controlling linguistic style aspects in neural language generation. In: Proceedings of the workshop on stylistic variation, association for computational linguistics, Copenhagen, Denmark, pp 94–104, https://doi.org/10.18653/v1/W17-4912
Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, PMLR, proceedings of machine learning research, vol 70, pp 1126–1135, http://proceedings.mlr.press/v70/finn17a.html
Fung P, Dey A, Siddique FB, Lin R, Yang Y, Bertero D, Wan Y, Chan RHY, Wu CS (2016) Zara: A virtual interactive dialogue system incorporating emotion, sentiment and personality recognition. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations, The COLING 2016 Organizing Committee, Osaka, Japan, pp 278–281, https://aclanthology.org/C16-2058
Galley M, Brockett C, Sordoni A, Ji Y, Auli M, Quirk C, Mitchell M, Gao J, Dolan B (2015) deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: short papers), association for computational linguistics, Beijing, China, pp 445–450, https://doi.org/10.3115/v1/P15-2073
Gan Z, Cheng Y, Kholy A, Li L, Liu J, Gao J (2019) Multi-step reasoning via recurrent dual attention for visual dialog. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 6463–6474, https://doi.org/10.18653/v1/P19-1648
Gan Z, Chen Y, Li L, Zhu C, Cheng Y, Liu J (2020) Large-scale adversarial training for vision-and-language representation learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, https://proceedings.neurips.cc/paper/2020/hash/49562478de4c54fafd4ec46fdb297de5-Abstract.html
Gangadharaiah R, Narayanaswamy B (2020) Recursive template-based frame generation for task oriented dialog. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2059–2064, https://doi.org/10.18653/v1/2020.acl-main.186
Gao J, Galley M, Li L (2018) Neural approaches to conversational AI. In: Collins-Thompson K, Mei Q, Davison BD, Liu Y, Yilmaz E (eds) The 41st international ACM SIGIR conference on research & development in information retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08–12, 2018, ACM, pp 1371–1374, https://doi.org/10.1145/3209978.3210183
Gao S, Zhang Y, Ou Z, Yu Z (2020a) Paraphrase augmented task-oriented dialog generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 639–649, https://doi.org/10.18653/v1/2020.acl-main.60
Gao X, Zhang Y, Lee S, Galley M, Brockett C, Gao J, Dolan B (2019) Structuring latent spaces for stylized response generation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), association for computational linguistics, Hong Kong, China, pp 1814–1823, https://doi.org/10.18653/v1/D19-1190
Gao X, Zhang Y, Galley M, Brockett C, Dolan B (2020b) Dialogue response ranking training with large-scale human feedback data. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 386–395, https://doi.org/10.18653/v1/2020.emnlp-main.28
Gao Y, Wu CS, Joty S, Xiong C, Socher R, King I, Lyu M, Hoi SC (2020c) Explicit memory tracker with coarse-to-fine reasoning for conversational machine reading. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 935–945, https://doi.org/10.18653/v1/2020.acl-main.88
Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, PMLR, proceedings of machine learning research, vol 70, pp 1243–1252, http://proceedings.mlr.press/v70/gehring17a.html

Ghazvininejad M, Brockett C, Chang M, Dolan B, Gao J, Yih W, Galley M (2018) A knowledge-grounded neural conversation model. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, pp 5110–5117, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16710

Gliwa B, Mochol I, Biesek M, Wawer A (2019) SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In: Proceedings of the 2nd workshop on new frontiers in summarization, association for computational linguistics, Hong Kong, China, pp 70–79, https://doi.org/10.18653/v1/D19-5409

Goddeau D, Meng H, Polifroni J, Seneff S, Busayapongchai S (1996) A form-based dialogue manager for spoken language applications. In: Proceeding of fourth international conference on spoken language processing. ICSLP'96, IEEE, vol 2, pp 701–704

Golovanov S, Kurbanov R, Nikolenko S, Truskovskyi K, Tselousov A, Wolf T (2019) Large-scale transfer learning for natural language generation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 6053–6058, https://doi.org/10.18653/v1/P19-1608

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv:1406.2661

Gopalakrishnan K, Hedayatnia B, Chen Q, Gottardi A, Kwatra S, Venkatesh A, Gabriel R, Hakkani-Tür D (2019) Topical-chat: Towards knowledge-grounded open-domain conversations. In: Kubin G, Kacic Z (eds) Interspeech 2019, 20th annual conference of the international speech communication association, Graz, Austria, 15–19 September 2019, ISCA, pp 1891–1895, https://doi.org/10.21437/Interspeech.2019-3079

Gordon-Hall G, Gorinski PJ, Cohen SB (2020) Learning dialog policies from weak demonstrations. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1394–1405, https://doi.org/10.18653/v1/2020.acl-main.129

Graves A, Wayne G, Danihelka I (2014) Neural turing machines

Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, Colmenarejo SG, Grefenstette E, Ramalho T, Agapiou J et al (2016) Hybrid computing using a neural network with dynamic external memory. Nature 538(7626):471–476

Gruber N, Jockisch A (2020) Are gru cells more specific and lstm cells more sensitive in motive classification of text? Front Artif Intell 3(40):1–6

Gu J, Lu Z, Li H, Li VO (2016) Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Berlin, Germany, pp 1631–1640, https://doi.org/10.18653/v1/P16-1154

Guo Q, Qiu X, Liu P, Shao Y, Xue X, Zhang Z (2019) Star-transformer. 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), ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, Minneapolis, Minnesota, pp 1315–1325, https://doi.org/10.18653/v1/N19-1133

Guo X, Yu M, Gao Y, Gan C, Campbell M, Chang S (2020) Interactive fiction game playing as multi-paragraph reading comprehension with reinforcement learning. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 7755–7765, https://doi.org/10.18653/v1/2020.emnlp-main.624

Gür I, Hakkani-Tür D, Tür G, Shah P (2018) User modeling for task oriented dialogues. In: 2018 IEEE spoken language technology workshop (SLT), IEEE, pp 900–906

Haber J, Baumgärtner T, Takmaz E, Gelderloos L, Bruni E, Fernández R (2019) The PhotoBook dataset: Building common ground through visually-grounded dialogue. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 1895–1910, https://doi.org/10.18653/v1/P19-1184

Hahn M, Krantz J, Batra D, Parikh D, Rehg J, Lee S, Anderson P (2020) Where are you? Localization from embodied dialog. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 806–822, https://doi.org/10.18653/v1/2020.emnlp-main.59

Hakkani-Tür D, Tür G, Celikyilmaz A, Chen Y, Gao J, Deng L, Wang Y (2016) Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM. In: Morgan N (ed) Interspeech 2016, 17th annual

conference of the international speech communication association, San Francisco, CA, USA, September 8–12, 2016, ISCA, pp 715–719, https://doi.org/10.21437/Interspeech.2016-402
Ham D, Lee JG, Jang Y, Kim KE (2020) End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 583–592, https://doi.org/10.18653/v1/2020.acl-main.54
Han M, Kang M, Jung H, Hwang SJ (2019) Episodic memory reader: Learning what to remember for question answering from streaming data. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 4407–4417, https://doi.org/10.18653/v1/P19-1434
Hancock B, Bordes A, Mazare PE, Weston J (2019) Learning from dialogue after deployment: Feed yourself, chatbot! In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3667–3684, https://doi.org/10.18653/v1/P19-1358
Hashemi HB, Asiaee A, Kraft R (2016) Query intent detection using convolutional neural networks. In: International conference on web search and data mining, workshop on query understanding
He H, Balakrishnan A, Eric M, Liang P (2017) Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Vancouver, Canada, pp 1766–1776, https://doi.org/10.18653/v1/P17-1162
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society, pp 770–778, https://doi.org/10.1109/CVPR.2016.90
He T, Glass J (2020) Negative training for neural dialogue response generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2044–2058, https://doi.org/10.18653/v1/2020.acl-main.185
He W, Yang M, Yan R, Li C, Shen Y, Xu R (2020a) Amalgamating knowledge from two teachers for task-oriented dialogue system with adversarial training. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3498–3507, https://doi.org/10.18653/v1/2020.emnlp-main.281
He X, Chen S, Ju Z, Dong X, Fang H, Wang S, Yang Y, Zeng J, Zhang R, Zhang R, et al. (2020b) Meddialog: Two large-scale medical dialogue datasets
Henderson J, Lemon O, Georgila K (2008) Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. Comput Linguist 34(4):487–511. https://doi.org/10.1162/coli.2008.07-028-R2-05-82
Henderson M (2015) Machine learning for dialog state tracking: A review. In: Proceedings of the first international workshop on machine learning in spoken language processing
Henderson M, Thomson B, Young S (2013) Deep neural network approach for the dialog state tracking challenge. In: Proceedings of the SIGDIAL 2013 conference, association for computational linguistics, Metz, France, pp 467–471, https://aclanthology.org/W13-4073
Henderson M, Thomson B, Williams JD (2014a) The second dialog state tracking challenge. In: Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL), association for computational linguistics, Philadelphia, PA, U.S.A., pp 263–272, https://doi.org/10.3115/v1/W14-4337
Henderson M, Thomson B, Williams JD (2014b) The third dialog state tracking challenge. In: 2014 IEEE spoken language technology workshop (SLT), IEEE, pp 324–329
Henderson M, Budzianowski P, Casanueva I, Coope S, Gerz D, Kumar G, Mrkšić N, Spithourakis G, Su PH, Vulić I, Wen TH (2019a) A repository of conversational datasets. In: Proceedings of the first workshop on NLP for conversational AI, association for computational linguistics, Florence, Italy, pp 1–10, https://doi.org/10.18653/v1/W19-4101
Henderson M, Vulić I, Gerz D, Casanueva I, Budzianowski P, Coope S, Spithourakis G, Wen TH, Mrkšić N, Su PH (2019b) Training neural response selection for task-oriented dialogue systems. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5392–5404, https://doi.org/10.18653/v1/P19-1536
Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
Hochreiter S, Bengio Y, Frasconi P, Schmidhuber J, et al. (2001) Gradient flow in recurrent nets: the difficulty of learning long-term dependencies
Hokamp C, Liu Q (2017) Lexically constrained decoding for sequence generation using grid beam search. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume

1: long papers), association for computational linguistics, Vancouver, Canada, pp 1535–1546, https://doi.org/10.18653/v1/P17-1141

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79(8):2554–2558

Hosseini-Asl E, McCann B, Wu C, Yavuz S, Socher R (2020) A simple language model for task-oriented dialogue. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, https://proceedings.neurips.cc/paper/2020/hash/e946209592563be0f01c844ab2170f0c-Abstract.html

Hu J, Yang Y, Chen C, He L, Yu Z (2020) SAS: Dialogue state tracking via slot attention and slot information sharing. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 6366–6375, https://doi.org/10.18653/v1/2020.acl-main.567

Hu JE, Rudinger R, Post M, Durme BV (2019) PARABANK: monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In: The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press, pp 6521–6528, https://doi.org/10.1609/aaai.v33i01.33016521

Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Toward controlled generation of text. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, PMLR, proceedings of machine learning research, vol 70, pp 1587–1596, http://proceedings.mlr.press/v70/hu17e.html

Hua X, Wang L (2019) Sentence-level content planning and style specification for neural text generation. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), association for computational linguistics, Hong Kong, China, pp 591–602, https://doi.org/10.18653/v1/D19-1055

Hua Y, Li YF, Haffari G, Qi G, Wu T (2020) Few-shot complex knowledge base question answering via meta reinforcement learning. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 5827–5837, https://doi.org/10.18653/v1/2020.emnlp-main.469

Huang L, Ye Z, Qin J, Lin L, Liang X (2020a) GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 9230–9240, https://doi.org/10.18653/v1/2020.emnlp-main.742

Huang X, Jiang J, Zhao D, Feng Y, Hong Y (2018) Natural language processing and Chinese computing: 6th CCF international conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings, vol 10619. Springer

Huang X, Qi J, Sun Y, Zhang R (2020b) Semi-supervised dialogue policy learning via stochastic reward estimation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 660–670, https://doi.org/10.18653/v1/2020.acl-main.62

Huang Y, Feng J, Hu M, Wu X, Du X, Ma S (2020c) Meta-reinforced multi-domain state generator for dialogue systems. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 7109–7118, https://doi.org/10.18653/v1/2020.acl-main.636

Huang Z, Zeng Z, Liu B, Fu D, Fu J (2020d) Pixel-bert: aligning image pixels with text by deep multi-modal transformers. https://arxiv.org/abs/2004.00849

Jaderberg M, Mnih V, Czarnecki WM, Schaul T, Leibo JZ, Silver D, Kavukcuoglu K (2017) Reinforcement learning with unsupervised auxiliary tasks. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings, OpenReview.net, https://openreview.net/forum?id=SJ6yPD5xg

Jang Y, Song Y, Yu Y, Kim Y, Kim G (2017) TGIF-QA: toward spatio-temporal reasoning in visual question answering. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, pp 1359–1367, https://doi.org/10.1109/CVPR.2017.149

Jaques N, Shen JH, Ghandeharioun A, Ferguson C, Lapedriza A, Jones N, Gu S, Picard R (2020) Human-centric dialog training via offline reinforcement learning. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3985–4003, https://doi.org/10.18653/v1/2020.emnlp-main.327

Ji C, Zhou X, Zhang Y, Liu X, Sun C, Zhu C, Zhao T (2020) Cross copy network for dialogue generation. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 1900–1910, https://doi.org/10.18653/v1/2020.emnlp-main.149

Ji G, He S, Xu L, Liu K, Zhao J (2015) Knowledge graph embedding via dynamic mapping matrix. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers), association for computational linguistics, Beijing, China, pp 687–696, https://doi.org/10.3115/v1/P15-1067

Ji S, Pan S, Cambria E, Marttinen P, Yu PS (2022) A survey on knowledge graphs: Representation, acquisition and applications. IEEE Trans Neural Netw Learn Syst 33(10):1–8

Jia Q, Liu Y, Ren S, Zhu K, Tang H (2020) Multi-turn response selection using dialogue dependency relations. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 1911–1920, https://doi.org/10.18653/v1/2020.emnlp-main.150

Jordan M (1986) Serial order: a parallel distributed processing approach. Technical report, June 1985–March 1986. Tech. rep., California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science

Jung J, Son B, Lyu S (2020) 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), association for computational linguistics, online, pp 3484–3497, https://doi.org/10.18653/v1/2020.emnlp-main.280

Jurafsky D (1997) Switchboard swbd-damsl shallow-discourse-function annotation coders manual. Institute of Cognitive Science Technical Report

K M A, Basu Roy Chowdhury S, Dukkipati A (2018) Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. In: Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long papers), association for computational linguistics, New Orleans, Louisiana, pp 313–322, https://doi.org/10.18653/v1/N18-1029

Kale M, Rastogi A (2020) Template guided text generation for task-oriented dialogue. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 6505–6520, https://doi.org/10.18653/v1/2020.emnlp-main.527

Kamezawa H, Nishida N, Shimizu N, Miyazaki T, Nakayama H (2020) A visually-grounded first-person dialogue dataset with verbal and non-verbal responses. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3299–3310, https://doi.org/10.18653/v1/2020.emnlp-main.267

Kannan A, Vinyals O (2017) Adversarial evaluation of dialogue models. https://arxiv.org/abs/1701.08198

Keskar NS, McCann B, Varshney LR, Xiong C, Socher R (2019) Ctrl: A conditional transformer language model for controllable generation. https://arxiv.org/abs/1909.05858

Kim A, Song HJ, Park SB, et al. (2018) A two-step neural dialog state tracker for task-oriented dialog processing. Computational intelligence and neuroscience 2018

Kim H, Kim B, Kim G (2020a) Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 904–916, https://doi.org/10.18653/v1/2020.emnlp-main.65

Kim S, D'Haro LF, Banchs RE, Williams JD, Henderson M, Yoshino K (2016) The fifth dialog state tracking challenge. In: 2016 IEEE Spoken Language Technology Workshop (SLT), IEEE, pp 511–517

Kim S, D'Haro LF, Banchs RE, Williams JD, Henderson M (2017) The fourth dialog state tracking challenge. In: Dialogues with social robots. Springer, pp 435–449

Kim S, Yang S, Kim G, Lee SW (2020b) Efficient dialogue state tracking by selectively overwriting memory. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 567–582, https://doi.org/10.18653/v1/2020.acl-main.53

Ko WJ, Ray A, Shen Y, Jin H (2020) Generating dialogue responses from a semantic latent space. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 4339–4349, https://doi.org/10.18653/v1/2020.emnlp-main.352

Konda VR, Tsitsiklis JN (2000) Actor-critic algorithms. In: Advances in neural information processing systems, Citeseer, pp 1008–1014

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural

information processing systems 25: 26th annual conference on neural information processing systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, United States, pp 1106–1114, https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Kummerfeld JK, Gouravajhala SR, Peper JJ, Athreya V, Gunasekara C, Ganhotra J, Patel SS, Polymenakos LC, Lasecki W (2019) A large-scale corpus for conversation disentanglement. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3846–3856, https://doi.org/10.18653/v1/P19-1374

Kundu S, Lin Q, Ng HT (2020) Learning to identify follow-up questions in conversational question answering. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 959–968 https://doi.org/10.18653/v1/2020.acl-main.90

Kurach K, Andrychowicz M, Sutskever I (2016) Neural random-access machines. In: Bengio Y, LeCun Y (eds) 4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, conference track proceedings, http://arxiv.org/abs/1511.06392

Larson S, Mahendran A, Peper JJ, Clarke C, Lee A, Hill P, Kummerfeld JK, Leach K, Laurenzano MA, Tang L, Mars J (2019) An evaluation dataset for intent classification and out-of-scope prediction. 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), association for computational linguistics, Hong Kong, China, pp 1311–1316, https://doi.org/10.18653/v1/D19-1131

Le H, Hoi SC (2020) Video-grounded dialogues with pretrained generation language models. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 5842–5848, https://doi.org/10.18653/v1/2020.acl-main.518

Le H, Sahoo D, Chen N, Hoi S (2019) Multimodal transformer networks for end-to-end video-grounded dialogue systems. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5612–5623, https://doi.org/10.18653/v1/P19-1564

Le H, Sahoo D, Chen N, Hoi SC (2020a) BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 1846–1859, https://doi.org/10.18653/v1/2020.emnlp-main.145

Le H, Sahoo D, Liu C, Chen N, Hoi SC (2020b) UniConv: a unified conversational neural architecture for multi-domain task-oriented dialogues. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 1860–1877, https://doi.org/10.18653/v1/2020.emnlp-main.146

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

Lee JY, Dernoncourt F (2016) Sequential short-text classification with recurrent and convolutional neural networks. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, association for computational linguistics, San Diego, California, pp 515–520, https://doi.org/10.18653/v1/N16-1062

Lee S (2013) Structured discriminative model for dialog state tracking. In: Proceedings of the SIGDIAL 2013 conference, association for computational linguistics, Metz, France, pp 442–451, https://aclanthology.org/W13-4069

Lee S, Eskenazi M (2013) Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In: Proceedings of the SIGDIAL 2013 conference, association for computational linguistics, Metz, France, pp 414–422, https://aclanthology.org/W13-4066

Lee S, Jha R (2019) Zero-shot adaptive transfer for conversational language understanding. In: The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, The Ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press, pp 6642–6649, https://doi.org/10.1609/aaai.v33i01.33016642

Lee S, Schulz H, Atkinson A, Gao J, Suleman K, El Asri L, Adada M, Huang M, Sharma S, Tay W et al (2019) Multi-domain task-completion dialog challenge. Dialog Syst Technol Chall 8:9

Lei W, Jin X, Kan MY, Ren Z, He X, Yin D (2018) Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Melbourne, Australia, pp 1437–1447, https://doi.org/10.18653/v1/P18-1133

Lemon O, Pietquin O (2007) Machine learning for spoken dialogue systems. In: Eighth annual conference of the international speech communication association

Li G, Duan N, Fang Y, Gong M, Jiang D (2020a) Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: The thirty-fourth aaai conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, AAAI Press, pp 11336–11344, https://aaai.org/ojs/index.php/AAAI/article/view/6795

Li J, Galley M, Brockett C, Gao J, Dolan B (2016a) 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, association for computational linguistics, San Diego, California, pp 110–119, https://doi.org/10.18653/v1/N16-1014

Li J, Monroe W, Jurafsky D (2016b) A simple, fast diverse decoding algorithm for neural generation. https://arxiv.org/abs/1611.08562

Li J, Monroe W, Ritter A, Jurafsky D, Galley M, Gao J (2016c) Deep reinforcement learning for dialogue generation. In: Proceedings of the 2016 conference on empirical methods in natural language processing, association for computational linguistics, Austin, Texas, pp 1192–1202, https://doi.org/10.18653/v1/D16-1127

Li J, Miller AH, Chopra S, Ranzato M, Weston J (2017a) Dialogue learning with human-in-the-loop. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24-26, 2017, conference track proceedings, OpenReview.net, https://openreview.net/forum?id=HJgXCV9xx

Li J, Miller AH, Chopra S, Ranzato M, Weston J (2017b) Learning through dialogue interactions by asking questions. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings, OpenReview.net, https://openreview.net/forum?id=rkE8pVcle

Li J, Monroe W, Shi T, Jean S, Ritter A, Jurafsky D (2017c) Adversarial learning for neural dialogue generation. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, Copenhagen, Denmark, pp 2157–2169, https://doi.org/10.18653/v1/D17-1230

Li L, Xu C, Wu W, Zhao Y, Zhao X, Tao C (2020b) Zero-resource knowledge-grounded dialogue generation. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual, https://proceedings.neurips.cc/paper/2020/hash/609c5e5089a9aa967232aba2a4d03114-Abstract.html

Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW (2019a) Visualbert: A simple and performant baseline for vision and language. https://arxiv.org/abs/1908.03557

Li M, Roller S, Kulikov I, Welleck S, Boureau YL, Cho K, Weston J (2020c) Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, Online, pp 4715–4728, https://doi.org/10.18653/v1/2020.acl-main.428

Li W, Shao W, Ji S, Cambria E (2022) Bieru: bidirectional emotional recurrent unit for conversational sentiment analysis. Neurocomputing 467:73–82

Li X, Lipton ZC, Dhingra B, Li L, Gao J, Chen YN (2016d) A user simulator for task-completion dialogues. https://arxiv.org/abs/1612.05688

Li X, Chen YN, Li L, Gao J, Celikyilmaz A (2017d) End-to-end task-completion neural dialogue systems. In: Proceedings of the eighth international joint conference on natural language processing (volume 1: long papers), Asian federation of natural language processing, Taipei, Taiwan, pp 733–743, https://aclanthology.org/I17-1074

Li X, Wang Y, Sun S, Panda S, Liu J, Gao J (2018) Microsoft dialogue challenge: building end-to-end task-completion dialogue systems. https://arxiv.org/abs/1807.11125

Li X, Yin F, Sun Z, Li X, Yuan A, Chai D, Zhou M, Li J (2019b) Entity-relation extraction as multi-turn question answering. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 1340–1350, https://doi.org/10.18653/v1/P19-1129

Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, et al. (2020d) Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European conference on computer vision, Springer, pp 121–137

Li Y (2017) Deep reinforcement learning: an overview. https://arxiv.org/abs/1701.07274

Li Y, Su H, Shen X, Li W, Cao Z, Niu S (2017e) DailyDialog: A manually labelled multi-turn dialogue dataset. In: Proceedings of the eighth international joint conference on natural language processing

(volume 1: long papers), Asian federation of natural language processing, Taipei, Taiwan, pp 986–995, https://aclanthology.org/I17-1099

Li Y, Yao K, Qin L, Che W, Li X, Liu T (2020e) Slot-consistent NLG for task-oriented dialogue systems with iterative rectification network. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 97–106, https://doi.org/10.18653/v1/2020.acl-main.10

Li Z, Niu C, Meng F, Feng Y, Li Q, Zhou J (2019c) Incremental transformer with deliberation decoder for document grounded conversations. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 12–21, https://doi.org/10.18653/v1/P19-1002

Liang W, Zou J, Yu Z (2020) Beyond user self-reported Likert scale ratings: a comparison model for automatic dialog evaluation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1363–1374, https://doi.org/10.18653/v1/2020.acl-main.126

Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: Text summarization branches out, association for computational linguistics, Barcelona, Spain, pp 74–81, https://aclanthology.org/W04-1013

Lin LJ (1992) Self-improving reactive agents based on reinforcement learning, planning and teaching. Mach Learn 8(3–4):293–321

Lin T, Wang Y, Liu X, Qiu X (2021) A survey of transformers. https://arxiv.org/abs/2106.04554

Lin X, Joty S, Jwalapuram P, Bari MS (2019) A unified linear-time framework for sentence-level discourse parsing. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 4190–4200, https://doi.org/10.18653/v1/P19-1410

Lin X, Jian W, He J, Wang T, Chu W (2020a) Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 41–52, https://doi.org/10.18653/v1/2020.acl-main.6

Lin Y, Liu Z, Sun M, Liu Y, Zhu X (2015) Learning entity and relation embeddings for knowledge graph completion. In: Bonet B, Koenig S (eds) Proceedings of the twenty-ninth AAAI conference on artificial intelligence, january 25–30, 2015, Austin, Texas, USA, AAAI Press, pp 2181–2187, http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9571

Lin Z, Cai D, Wang Y, Liu X, Zheng H, Shi S (2020b) The world is not binary: Learning to rank with grayscale data for dialogue response selection. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 9220–9229, https://doi.org/10.18653/v1/2020.emnlp-main.741

Lin Z, Madotto A, Winata GI, Fung P (2020c) MinTL: Minimalist transfer learning for task-oriented dialogue systems. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3391–3405, https://doi.org/10.18653/v1/2020.emnlp-main.273

Lipton ZC, Berkowitz J, Elkan C (2015) A critical review of recurrent neural networks for sequence learning. https://arxiv.org/abs/1506.00019

Lison P, Bibauw S (2017) Not all dialogues are created equal: Instance weighting for neural conversational models. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics, Saarbrücken, Germany, pp 384–394, https://doi.org/10.18653/v1/W17-5546

Liu B, Lane I (2017) Iterative policy learning in end-to-end trainable task-oriented neural dialog models. In: 2017 IEEE automatic speech recognition and understanding workshop (ASRU), IEEE, pp 482–489

Liu B, Lane IR (2016) Attention-based recurrent neural network models for joint intent detection and slot filling. In: Morgan N (ed) Interspeech 2016, 17th annual conference of the international speech communication association, San Francisco, CA, USA, September 8–12, 2016, ISCA, pp 685–689, https://doi.org/10.21437/Interspeech.2016-1352

Liu C, He S, Liu K, Zhao J (2019) Vocabulary pyramid network: Multi-pass encoding and decoding with multi-level vocabularies for response generation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3774–3783, https://doi.org/10.18653/v1/P19-1367

Liu CW, Lowe R, Serban I, Noseworthy M, Charlin L, Pineau J (2016) How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: Proceedings of the 2016 conference on empirical methods in natural language processing, association for computational linguistics, Austin, Texas, pp 2122–2132, https://doi.org/10.18653/v1/D16-1230

Liu H, Wang W, Wang Y, Liu H, Liu Z, Tang J (2020a) Mitigating gender bias for neural dialogue generation with adversarial learning. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 893–903, https://doi.org/10.18653/v1/2020.emnlp-main.64

Liu Q, Chen Y, Chen B, Lou JG, Chen Z, Zhou B, Zhang D (2020b) You impress me: dialogue generation via mutual persona perception. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1417–1427, https://doi.org/10.18653/v1/2020.acl-main.131

Liu Y, Lapata M (2018) Learning structured text representations. Trans Assoc Comput Linguist 6:63–75

Liu Z, Wang H, Niu ZY, Wu H, Che W, Liu T (2020c) Towards conversational recommendation over multi-type dialogs. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1036–1049, https://doi.org/10.18653/v1/2020.acl-main.98

Lowe R, Pow N, Serban I, Pineau J (2015) The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In: Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue, association for computational linguistics, Prague, Czech Republic, pp 285–294, https://doi.org/10.18653/v1/W15-4640

Lowe R, Noseworthy M, Serban IV, Angelard-Gontier N, Bengio Y, Pineau J (2017) Towards an automatic Turing test: Learning to evaluate dialogue responses. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Vancouver, Canada, pp 1116–1126, https://doi.org/10.18653/v1/P17-1103

Lu J, Batra D, Parikh D, Lee S (2019a) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp 13–23, https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html

Lu J, Zhang C, Xie Z, Ling G, Zhou TC, Xu Z (2019b) Constructing interpretive spatio-temporal features for multi-turn responses selection. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 44–50, https://doi.org/10.18653/v1/P19-1006

Lu J, Goswami V, Rohrbach M, Parikh D, Lee S (2020) 12-in-1: Multi-task vision and language representation learning. In: 2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020, IEEE, pp 10434–10443, https://doi.org/10.1109/CVPR42600.2020.01045

Lubis N, Sakti S, Yoshino K, Nakamura S (2018) Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, pp 5293–5300, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16317

Ma MD, Bowden K, Wu J, Cui W, Walker M (2019) Implicit discourse relation identification for open-domain dialogues. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 666–672, https://doi.org/10.18653/v1/P19-1065

Ma W, Cui Y, Liu T, Wang D, Wang S, Hu G (2020a) Conversational Word Embedding for Retrieval-Based Dialog System. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1375–1380, https://doi.org/10.18653/v1/2020.acl-main.127

Ma Y, Nguyen KL, Xing FZ, Cambria E (2020) A survey on empathetic dialogue systems. Inf Fusion 64:50–70

Madotto A, Lin Z, Wu CS, Fung P (2019) Personalizing dialogue agents via meta-learning. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5454–5459, https://doi.org/10.18653/v1/P19-1542

Majumder BP, Jhamtani H, Berg-Kirkpatrick T, McAuley J (2020a) Like hiking? You probably enjoy nature: Persona-grounded dialog with commonsense expansions. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 9194–9206, https://doi.org/10.18653/v1/2020.emnlp-main.739

Majumder BP, Li S, Ni J, McAuley J (2020b) Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 8129–8141, https://doi.org/10.18653/v1/2020.emnlp-main.653

Mallios S, Bourbakis N (2016) A survey on human machine dialogue systems. In: 2016 7th international conference on information, intelligence, systems & applications (IISA), IEEE, pp 1–7

Manuvirakurike R, Brixey J, Bui T, Chang W, Artstein R, Georgila K (2018) DialEdit: Annotations for spoken conversational image editing. In: Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 1–9, https://aclanthology.org/W18-4701

Mao HH, Li S, McAuley JJ, Cottrell GW (2020) Speech recognition and multi-speaker diarization of long conversations. In: Meng H, Xu B, Zheng TF (eds) Interspeech 2020, 21st Annual conference of the international speech communication association, virtual event, Shanghai, China, 25–29 October 2020, ISCA, pp 691–695, https://doi.org/10.21437/Interspeech.2020-3039

Mehri S, Eskenazi M (2020) USR: An unsupervised and reference free evaluation metric for dialog generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 681–707, https://doi.org/10.18653/v1/2020.acl-main.64

Mehri S, Razumovskaia E, Zhao T, Eskenazi M (2019) Pretraining methods for dialog context representation learning. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3836–3845, https://doi.org/10.18653/v1/P19-1373

Mesgar M, Bücker S, Gurevych I (2020) Dialogue coherence assessment without explicit dialogue act labels. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1439–1450, https://doi.org/10.18653/v1/2020.acl-main.133

Mesnil G, He X, Deng L, Bengio Y (2013) Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In: Interspeech, pp 3771–3775

Mesnil G, Dauphin Y, Yao K, Bengio Y, Deng L, Hakkani-Tur D, He X, Heck L, Tur G, Yu D et al (2014) Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Trans Audio Speech Language Process 23(3):530–539

Miao N, Zhou H, Mou L, Yan R, Li L (2019) CGMH: constrained sentence generation by metropolis-hastings sampling. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, pp 6834–6842, https://doi.org/10.1609/aaai.v33i01.33016834

Miech A, Alayrac J, Smaira L, Laptev I, Sivic J, Zisserman A (2020) End-to-end learning of visual representations from uncurated instructional videos. In: 2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020, IEEE, pp 9876–9886, https://doi.org/10.1109/CVPR42600.2020.00990

Miller A, Fisch A, Dodge J, Karimi AH, Bordes A, Weston J (2016) Key-value memory networks for directly reading documents. In: Proceedings of the 2016 conference on empirical methods in natural language processing, association for computational linguistics, Austin, Texas, pp 1400–1409, https://doi.org/10.18653/v1/D16-1147

Miltsakaki E, Prasad R, Joshi A, Webber B (2004) The Penn Discourse Treebank. In: Proceedings of the fourth international conference on language resources and evaluation (LREC'04), European Language Resources Association (ELRA), Lisbon, Portugal, http://www.lrec-conf.org/proceedings/lrec2004/pdf/618.pdf

Mirowski P, Pascanu R, Viola F, Soyer H, Ballard A, Banino A, Denil M, Goroshin R, Sifre L, Kavukcuoglu K, Kumaran D, Hadsell R (2017) Learning to navigate in complex environments. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings, OpenReview.net, https://openreview.net/forum?id=SJMGPrcle

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533

Mnih V, Badia AP, Mirza M, Graves A, Lillicrap TP, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: Balcan M, Weinberger KQ (eds) Proceedings of the 33nd international conference on machine learning, ICML 2016, New York City, NY, USA, June

19–24, 2016, JMLR.org, JMLR Workshop and Conference Proceedings, vol 48, pp 1928–1937, http://proceedings.mlr.press/v48/mniha16.html

Mo K, Zhang Y, Li S, Li J, Yang Q (2018) Personalizing a dialogue system with transfer reinforcement learning. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, pp 5317–5324, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16104

Moghe N, Arora S, Banerjee S, Khapra MM (2018) Towards exploiting background knowledge for building conversation systems. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics, Brussels, Belgium, pp 2322–2332, https://doi.org/10.18653/v1/D18-1255

Mohammad S, Bravo-Marquez F, Salameh M, Kiritchenko S (2018) SemEval-2018 task 1: Affect in tweets. In: Proceedings of the 12th international workshop on semantic evaluation, association for computational linguistics, New Orleans, Louisiana, pp 1–17, https://doi.org/10.18653/v1/S18-1001

Moon S, Shah P, Kumar A, Subba R (2019) OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 845–854, https://doi.org/10.18653/v1/P19-1081

Mostafazadeh N, Brockett C, Dolan B, Galley M, Gao J, Spithourakis G, Vanderwende L (2017) Image-grounded conversations: multimodal context for natural question and response generation. In: Proceedings of the eighth international joint conference on natural language processing (volume 1: long papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, pp 462–472, https://aclanthology.org/I17-1047

Mrkšić N, Ó Séaghdha D, Thomson B, Gašić M, Su PH, Vandyke D, Wen TH, Young S (2015) Multi-domain dialog state tracking using recurrent neural networks. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: short papers), association for computational linguistics, Beijing, China, pp 794–799, https://doi.org/10.3115/v1/P15-2130

Mrkšić N, Ó Séaghdha D, Wen TH, Thomson B, Young S (2017) Neural belief tracker: data-driven dialogue state tracking. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Vancouver, Canada, pp 1777–1788, https://doi.org/10.18653/v1/P17-1163

Nakov P, Màrquez L, Magdy W, Moschitti A, Glass J, Randeree B (2015) SemEval-2015 task 3: Answer selection in community question answering. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), association for computational linguistics, Denver, Colorado, pp 269–281, https://doi.org/10.18653/v1/S15-2047

Ni J, Pandelea V, Young T, Zhou H, Cambria E (2022) Hitkg: Towards goal-oriented conversations via multi-hierarchy learning. Proceedings of the AAAI conference on artificial intelligence 36:11112–11120

Nickel M, Rosasco L, Poggio TA (2016) Holographic embeddings of knowledge graphs. In: Schuurmans D, Wellman MP (eds) Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA, AAAI Press, pp 1955–1961, http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12484

Novikova J, Dušek O, Rieser V (2017) The E2E dataset: new challenges for end-to-end generation. In: Proceedings of the 18th annual sigdial meeting on discourse and dialogue, association for computational linguistics, Saarbrücken, Germany, pp 201–206, https://doi.org/10.18653/v1/W17-5525

Obuchowski A, Lew M (2020) Transformer-capsule model for intent detection (student abstract). In: The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, AAAI Press, pp 13885–13886, https://aaai.org/ojs/index.php/AAAI/article/view/7215

Oraby S, Harrison V, Ebrahimi A, Walker M (2019) Curate and generate: a corpus and method for joint control of semantics and style in neural NLG. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5938–5951, https://doi.org/10.18653/v1/P19-1596

Ouyang Y, Chen M, Dai X, Zhao Y, Huang S, Chen J (2020) Dialogue state tracking with explicit slot connection modeling. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 34–40, https://doi.org/10.18653/v1/2020.acl-main.5

Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19–24, 2015, IEEE, pp 5206–5210, https://doi.org/10.1109/ICASSP.2015.7178964

Pang B, Nijkamp E, Han W, Zhou L, Liu Y, Tu K (2020) Towards holistic and automatic evaluation of open-domain dialogue generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 3619–3629, https://doi.org/10.18653/v1/2020.acl-main.333

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, association for computational linguistics, Philadelphia, Pennsylvania, USA, pp 311–318, https://doi.org/10.3115/1073083.1073135

Parikh A, Täckström O, Das D, Uszkoreit J (2016) A decomposable attention model for natural language inference. In: Proceedings of the 2016 conference on empirical methods in natural language processing, association for computational linguistics, Austin, Texas, pp 2249–2255, https://doi.org/10.18653/v1/D16-1244

Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. In: Proceedings of the 30th international conference on machine learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013, JMLR.org, JMLR workshop and conference proceedings, vol 28, pp 1310–1318, http://proceedings.mlr.press/v28/pascanu13.html

Peng B, Li X, Li L, Gao J, Celikyilmaz A, Lee S, Wong KF (2017) Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, Copenhagen, Denmark, pp 2231–2240, https://doi.org/10.18653/v1/D17-1237

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers), association for computational linguistics, New Orleans, Louisiana, pp 2227–2237, https://doi.org/10.18653/v1/N18-1202

Pfau D, Vinyals O (2016) Connecting generative adversarial networks and actor-critic methods

Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R (2019) MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 527–536, https://doi.org/10.18653/v1/P19-1050

Powers DMW (2020) Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation

Puterman ML (2014) Markov decision processes: discrete stochastic dynamic programming. Wiley, Hoboken

Qi D, Su L, Song J, Cui E, Bharti T, Sacheti A (2020) Imagebert: cross-modal pre-training with large-scale weak-supervised image-text data. https://arxiv.org/abs/2001.07966

Qian K, Yu Z (2019) Domain adaptive dialog generation via meta learning. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 2639–2649, https://doi.org/10.18653/v1/P19-1253

Qin L, Che W, Li Y, Wen H, Liu T (2019) A stack-propagation framework with token-level intent detection for spoken language understanding. 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), association for computational linguistics, Hong Kong, China, pp 2078–2087, https://doi.org/10.18653/v1/D19-1214

Qin L, Xu X, Che W, Zhang Y, Liu T (2020) Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 6344–6354, https://doi.org/10.18653/v1/2020.acl-main.565

Qiu L, Li J, Bi W, Zhao D, Yan R (2019) Are training samples correlated? Learning to generate dialogue responses with multiple references. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3826–3835, https://doi.org/10.18653/v1/P19-1372

Qiu L, Zhao Y, Shi W, Liang Y, Shi F, Yuan T, Yu Z, Zhu SC (2020) Structured attention for unsupervised dialogue structure induction. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 1889–1899, https://doi.org/10.18653/v1/2020.emnlp-main.148

Qiu M, Li FL, Wang S, Gao X, Chen Y, Zhao W, Chen H, Huang J, Chu W (2017) AliMe chat: A sequence to sequence and rerank based chatbot engine. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers), association for computational linguistics, Vancouver, Canada, pp 498–503, https://doi.org/10.18653/v1/P17-2079

Quan J, Xiong D (2020) Modeling long context for task-oriented dialogue state generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 7119–7124, https://doi.org/10.18653/v1/2020.acl-main.637

Quan J, Zhang S, Cao Q, Li Z, Xiong D (2020) RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 930–940, https://doi.org/10.18653/v1/2020.emnlp-main.67

Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: Unanswerable questions for SQuAD. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers), association for computational linguistics, Melbourne, Australia, pp 784–789, https://doi.org/10.18653/v1/P18-2124

Ram A, Prasad R, Khatri C, Venkatesh A, Gabriel R, Liu Q, Nunn J, Hedayatnia B, Cheng M, Nagar A, et al. (2018) Conversational ai: the science behind the alexa prize. https://arxiv.org/abs/1801.03604

Rameshkumar R, Bailey P (2020) Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 5121–5134, https://doi.org/10.18653/v1/2020.acl-main.459

Ramshaw L, Marcus M (1995) Text chunking using transformation-based learning. In: Third workshop on very large corpora, https://aclanthology.org/W95-0107

Rashkin H, Smith EM, Li M, Boureau YL (2019) Towards empathetic open-domain conversation models: A new benchmark and dataset. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5370–5381, https://doi.org/10.18653/v1/P19-1534

Rastogi A, Zang X, Sunkara S, Gupta R, Khaitan P (2020) Towards scalable multi-domain conversational agents: the schema-guided dialogue dataset. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, AAAI Press, pp 8689–8696, https://aaai.org/ojs/index.php/AAAI/article/view/6394

Ravuri S, Stolcke A (2015) Recurrent neural network and lstm models for lexical utterance classification. In: Sixteenth annual conference of the international speech communication association

Ravuri SV, Stolcke A (2016) A comparative study of recurrent neural network models for lexical domain classification. In: 2016 IEEE international conference on acoustics, speech and signal processing, ICASSP 2016, Shanghai, China, March 20–25, 2016, IEEE, pp 6075–6079, https://doi.org/10.1109/ICASSP.2016.7472844

Rawat W, Wang Z (2017) Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput 29(9):2352–2449

Reiter E (1994) Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In: Proceedings of the Seventh International Workshop on Natural Language Generation, https://aclanthology.org/W94-0319

Ren H, Xu W, Zhang Y, Yan Y (2013) Dialog state tracking using conditional random fields. In: Proceedings of the SIGDIAL 2013 conference, association for computational linguistics, Metz, France, pp 457–461, https://aclanthology.org/W13-4071

Ren L, Xie K, Chen L, Yu K (2018) Towards universal dialogue state tracking. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics, Brussels, Belgium, pp 2780–2786, https://doi.org/10.18653/v1/D18-1299

Ren P, Chen Z, Ren Z, Kanoulas E, Monz C, de Rijke M (2020) Conversations with search engines. https://arxiv.org/abs/2004.14162

Ritter A, Cherry C, Dolan WB (2011) Data-driven response generation in social media. In: Proceedings of the 2011 conference on empirical methods in natural language processing, association for computational linguistics, Edinburgh, Scotland, UK. pp 583–593, https://aclanthology.org/D11-1054

Rodriguez P, Crook P, Moon S, Wang Z (2020) Information seeking in the spirit of learning: a dataset for conversational curiosity. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 8153–8172, https://doi.org/10.18653/v1/2020.emnlp-main.655

Saha A, Khapra MM, Sankaranarayanan K (2018) Towards building large scale multimodal domain-aware conversation systems. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, pp 696–704, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17104

Saha T, Patra A, Saha S, Bhattacharyya P (2020) Towards emotion-aided multi-modal dialogue act classification. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 4361–4372, https://doi.org/10.18653/v1/2020.acl-main.402

Sankar C, Subramanian S, Pal C, Chandar S, Bengio Y (2019) Do neural dialog systems use the conversation history effectively? An empirical study. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 32–37, https://doi.org/10.18653/v1/P19-1004

Santhanam S, Shaikh S (2019) A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. https://arxiv.org/abs/1906.00500

Sarikaya R, Hinton GE, Ramabhadran B (2011) Deep belief nets for natural language call-routing. In: 2011 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5680–5683

Sarikaya R, Hinton GE, Deoras A (2014) Application of deep belief networks for natural language understanding. IEEE/ACM Trans Audio Speech Lang Process 22(4):778–784

Sato S, Akama R, Ouchi H, Suzuki J, Inui K (2020) Evaluating dialogue generation systems via response selection. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 593–599, https://doi.org/10.18653/v1/2020.acl-main.55

Schatzmann J, Young S (2009) The hidden agenda user simulation model. IEEE/ACM Trans Audio Speech Lang Process 17(4):733–747

Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681

See A, Roller S, Kiela D, Weston J (2019) What makes a good conversation? how controllable attributes affect human judgments. 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), association for computational linguistics, Minneapolis, Minnesota, pp 1702–1723, https://doi.org/10.18653/v1/N19-1170

Serban IV, Sordoni A, Bengio Y, Courville AC, Pineau J (2016) Building end-to-end dialogue systems using generative hierarchical neural network models. In: Schuurmans D, Wellman MP (eds) Proceedings of the thirtieth AAAI conference on artificial intelligence, February 12–17, 2016, Phoenix, Arizona, USA, AAAI Press, pp 3776–3784, http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957

Serban IV, Sankar C, Germain M, Zhang S, Lin Z, Subramanian S, Kim T, Pieper M, Chandar S, Ke NR, et al. (2017a) A deep reinforcement learning chatbot. https://arxiv.org/abs/1709.02349

Serban IV, Sordoni A, Lowe R, Charlin L, Pineau J, Courville AC, Bengio Y (2017b) A hierarchical latent variable encoder-decoder model for generating dialogues. In: Singh SP, Markovitch S (eds) Proceedings of the thirty-first AAAI conference on artificial intelligence, February 4–9, 2017, San Francisco, California, USA, AAAI Press, pp 3295–3301, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567

Serras M, Torres MI, del Pozo A (2019) Goal-conditioned user modeling for dialogue systems using stochastic bi-automata. In: ICPRAM, pp 128–134

Shah P, Hakkani-Tür D, Tür G, Rastogi A, Bapna A, Nayak N, Heck L (2018) Building a conversational agent overnight with dialogue self-play. https://arxiv.org/abs/1801.04871

Shan Y, Li Z, Zhang J, Meng F, Feng Y, Niu C, Zhou J (2020) A contextual hierarchical attention network with adaptive objective for dialogue state tracking. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 6322–6333, https://doi.org/10.18653/v1/2020.acl-main.563

Shang L, Lu Z, Li H (2015) Neural responding machine for short-text conversation. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers), association for computational linguistics, Beijing, China, pp 1577–1586, https://doi.org/10.3115/v1/P15-1152

Shao L, Gouws S, Britz D, Goldie A, Strope B, Kurzweil R (2017) Generating long and diverse responses with neural conversation models. https://arxiv.org/abs/1701.03185

Shao Y, Nakashole N (2020) ChartDialogs: Plotting from Natural Language Instructions. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 3559–3574, https://doi.org/10.18653/v1/2020.acl-main.328

Shen L, Feng Y, Zhan H (2019) Modeling semantic relationship in multi-turn conversations with hierarchical latent variables. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5497–5502, https://doi.org/10.18653/v1/P19-1549

Shi B, Weninger T (2017) Proje: Embedding projection for knowledge graph completion. In: Singh SP, Markovitch S (eds) Proceedings of the thirty-first AAAI conference on artificial intelligence, February 4–9, 2017, San Francisco, California, USA, AAAI Press, pp 1236–1242, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14279

Shuster K, Humeau S, Bordes A, Weston J (2020a) Image-chat: Engaging grounded conversations. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2414–2429, https://doi.org/10.18653/v1/2020.acl-main.219

Shuster K, Humeau S, Bordes A, Weston J (2020b) Image-chat: engaging grounded conversations. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2414–2429, https://doi.org/10.18653/v1/2020.acl-main.219

Shuster K, Ju D, Roller S, Dinan E, Boureau YL, Weston J (2020c) The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2453–2470, https://doi.org/10.18653/v1/2020.acl-main.222

Siddharthan A (2001) Ehud reiter and robert dale. Building natural language generation systems. Natural Lang Eng 7(3):271

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M et al (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484–489

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, http://arxiv.org/abs/1409.1556

Singh A, Goswami V, Parikh D (2020) Are we pretraining it right? Digging deeper into visio-linguistic pretraining. https://arxiv.org/abs/2004.08744

Singh S, Litman D, Kearns M, Walker M (2002) Optimizing dialogue management with reinforcement learning: experiments with the njfun system. J Artif Intell Res 16:105–133

Singla K, Chen Z, Atkins D, Narayanan S (2020) Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 3797–3803, https://doi.org/10.18653/v1/2020.acl-main.351

Sinha K, Parthasarathi P, Wang J, Lowe R, Hamilton WL, Pineau J (2020) Learning an unreferenced metric for online dialogue evaluation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2430–2441, https://doi.org/10.18653/v1/2020.acl-main.220

Smith EM, Williamson M, Shuster K, Weston J, Boureau YL (2020) Can you put it all together: Evaluating conversational agents' ability to blend skills. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2021–2030, https://doi.org/10.18653/v1/2020.acl-main.183

Socher R, Chen D, Manning CD, Ng AY (2013) Reasoning with neural tensor networks for knowledge base completion. In: Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5–8, 2013, Lake Tahoe, Nevada, United States, pp 926–934, https://proceedings.neurips.cc/paper/2013/hash/b337e84de8752b27eda3a12363109e80-Abstract.html

Song H, Wang Y, Zhang WN, Liu X, Liu T (2020a) Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 5821–5831, https://doi.org/10.18653/v1/2020.acl-main.516

Song H, Wang Y, Zhang WN, Zhao Z, Liu T, Liu X (2020b) Profile consistency identification for open-domain dialogue agents. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 6651–6662, https://doi.org/10.18653/v1/2020.emnlp-main.539

Song Y, Yan R, Li X, Zhao D, Zhang M (2016) Two are better than one: an ensemble of retrieval-and generation-based dialog systems

Song Z, Zheng X, Liu L, Xu M, Huang X (2019) Generating responses with a specific emotion in dialog. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3685–3695, https://doi.org/10.18653/v1/P19-1359

Sordoni A, Bengio Y, Vahabi H, Lioma C, Simonsen JG, Nie J (2015a) A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In: Bailey J, Moffat A, Aggarwal CC, de Rijke M, Kumar R, Murdock V, Sellis TK, Yu JX (eds) Proceedings of the 24th ACM international conference on information and knowledge management, CIKM 2015, Melbourne, VIC, Australia, October 19–23, 2015, ACM, pp 553–562, https://doi.org/10.1145/2806416.2806493

Sordoni A, Galley M, Auli M, Brockett C, Ji Y, Mitchell M, Nie JY, Gao J, Dolan B (2015b) A neural network approach to context-sensitive generation of conversational responses. In: Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies, association for computational linguistics, Denver, Colorado, pp 196–205, https://doi.org/10.3115/v1/N15-1020

Stasaski K, Yang GH, Hearst MA (2020) More diverse dialogue datasets via diversity-informed data collection. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 4958–4968, https://doi.org/10.18653/v1/2020.acl-main.446

Stent A, Marge M, Singhai M (2005) Evaluating evaluation methods for generation in the presence of variation. In: International conference on intelligent text processing and computational linguistics, Springer, pp 341–351

Su H, Shen X, Zhang R, Sun F, Hu P, Niu C, Zhou J (2019a) Improving multi-turn dialogue modelling with utterance ReWriter. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 22–31, https://doi.org/10.18653/v1/P19-1003

Su H, Shen X, Zhao S, Xiao Z, Hu P, Zhong R, Niu C, Zhou J (2020a) Diversifying dialogue generation with non-conversational text. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 7087–7097, https://doi.org/10.18653/v1/2020.acl-main.634

Su PH, Vandyke D, Gasic M, Kim D, Mrksic N, Wen TH, Young S (2015) Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. https://arxiv.org/abs/1508.03386

Su PH, Gasic M, Mrksic N, Rojas-Barahona L, Ultes S, Vandyke D, Wen TH, Young S (2016) Continuously learning neural dialogue management. https://arxiv.org/abs/1606.02689

Su SY, Huang CW, Chen YN (2019b) Dual supervised learning for natural language understanding and generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 5472–5477, https://doi.org/10.18653/v1/P19-1545

Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, Dai J (2020b) VL-BERT: pre-training of generic visual-linguistic representations. In: 8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, https://openreview.net/forum?id=SygXPaEYvH

Sukhbaatar S, Szlam A, Weston J, Fergus R (2015) End-to-end memory networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, pp 2440–2448, https://proceedings.neurips.cc/paper/2015/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html

Sun C, Baradel F, Murphy K, Schmid C (2019a) Learning video representations using contrastive bidirectional transformer. https://arxiv.org/abs/1906.05743

Sun C, Myers A, Vondrick C, Murphy K, Schmid C (2019b) Videobert: a joint model for video and language representation learning. In: 2019 IEEE/CVF international conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019, IEEE, pp 7463–7472, https://doi.org/10.1109/ICCV.2019.00756

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada, pp 3104–3112, https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3(1):9–44
Sutton RS, McAllester DA, Singh SP, Mansour Y et al (1999) Policy gradient methods for reinforcement learning with function approximation. NIPs, Citeseer 99:1057–1063
Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: IEEE conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, IEEE Computer Society, pp 1–9, https://doi.org/10.1109/CVPR.2015.7298594
Takanobu R, Liang R, Huang M (2020) Multi-agent task-oriented dialog policy learning with role-aware reward decomposition. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 625–638, https://doi.org/10.18653/v1/2020.acl-main.59
Takmaz E, Giulianelli M, Pezzelle S, Sinclair A, Fernández R (2020) Refer, reuse, reduce: generating subsequent references in visual and conversational contexts. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 4350–4368, https://doi.org/10.18653/v1/2020.emnlp-main.353
Tamar A, Levine S, Abbeel P, Wu Y, Thomas G (2016) Value iteration networks. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R (eds) Advances in neural information processing systems 29: annual conference on neural information processing systems 2016, December 5–10, 2016, Barcelona, Spain, pp 2146–2154, https://proceedings.neurips.cc/paper/2016/hash/c21002f464c5fc5bee3b98ced83963b8-Abstract.html
Tan H, Bansal M (2019) LXMERT: Learning cross-modality encoder representations from transformers. 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), Association for Computational Linguistics, Hong Kong, China, pp 5100–5111, https://doi.org/10.18653/v1/D19-1514
Tanana M, Hallgren KA, Imel ZE, Atkins DC, Srikumar V (2016) A comparison of natural language processing methods for automated coding of motivational interviewing. J Subst Abuse Treatment 65:43–50
Tang D, Qin B, Liu T (2015) Learning semantic representations of users and products for document level sentiment classification. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers), association for computational linguistics, Beijing, China, pp 1014–1023, https://doi.org/10.3115/v1/P15-1098
Tang J, Zhao T, Xiong C, Liang X, Xing E, Hu Z (2019) Target-guided open-domain conversation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5624–5634, https://doi.org/10.18653/v1/P19-1565
Tao C, Mou L, Zhao D, Yan R (2018) RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, pp 722–729, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16179
Tao C, Wu W, Xu C, Hu W, Zhao D, Yan R (2019) One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 1–11, https://doi.org/10.18653/v1/P19-1001
Tay Y, Wang S, Luu AT, Fu J, Phan MC, Yuan X, Rao J, Hui SC, Zhang A (2019) Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 4922–4931, https://doi.org/10.18653/v1/P19-1486
Tay Y, Dehghani M, Bahri D, Metzler D (2020) Efficient transformers: a survey. https://arxiv.org/abs/2009.06732
Theune M (2003) Natural language generation for dialogue: system survey. University of Twente, Centre for Telematics and Information Technology
Thomas M, Pang B, Lee L (2006) Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the 2006 conference on empirical methods in natural language processing, association for computational linguistics, Sydney, Australia, pp 327–335, https://aclanthology.org/W06-1639
Tian Z, Bi W, Li X, Zhang NL (2019) Learning to abstract for memory-augmented conversational response generation. In: Proceedings of the 57th annual meeting of the association for computational

linguistics, association for computational linguistics, Florence, Italy, pp 3816–3825, https://doi.org/10.18653/v1/P19-1371

Tiedemann J (2012) Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, pp 2214–2218, http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Tonelli S, Riccardi G, Prasad R, Joshi A (2010) Annotation of discourse relations for conversational spoken dialogs. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, http://www.lrec-conf.org/proceedings/lrec2010/pdf/184_Paper.pdf

Tran VK, Nguyen LM (2017) Semantic refinement gru-based neural language generation for spoken dialogue systems. In: International Conference of the Pacific Association for Computational Linguistics, Springer, pp 63–75

Tu G, Wen J, Liu C, Jiang D, Cambria E (2022) Context-and sentiment-aware networks for emotion recognition in conversation. IEEE Trans Artif Intell

Tur G, Hakkani-Tür D, Heck L (2010) What is left to be understood in atis? In: 2010 IEEE spoken language technology workshop, IEEE, pp 19–24

Tur G, Deng L, Hakkani-Tür D, He X (2012) Towards deeper understanding: deep convex networks for semantic utterance classification. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5045–5048

Ultes S, Rojas-Barahona LM, Su PH, Vandyke D, Kim D, Casanueva I, Budzianowski P, Mrkšić N, Wen TH, Gašić M, Young S (2017) PyDial: A multi-domain statistical dialogue system toolkit. In: Proceedings of ACL 2017, system demonstrations, association for computational linguistics, Vancouver, Canada, pp 73–78, https://aclanthology.org/P17-4013

Urbanek J, Fan A, Karamcheti S, Jain S, Humeau S, Dinan E, Rocktäschel T, Kiela D, Szlam A, Weston J (2019) Learning to speak and act in a fantasy text adventure game. 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), association for computational linguistics, Hong Kong, China, pp 673–683, https://doi.org/10.18653/v1/D19-1062

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 5998–6008, https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vijayakumar AK, Cogswell M, Selvaraju RR, Sun Q, Lee S, Crandall D, Batra D (2016) Diverse beam search: decoding diverse solutions from neural sequence models

Vinyals O, Le Q (2015) A neural conversational model. https://arxiv.org/abs/1506.05869

Vinyals O, Fortunato M, Jaitly N (2015) Pointer networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pp 2692–2700, https://proceedings.neurips.cc/paper/2015/hash/29921001f2f04bd3baee84a12e98098f-Abstract.html

Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 13(2):260–269

Vougiouklis P, Hare J, Simperl E (2016) A neural network approach for knowledge-driven response generation. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, The COLING 2016 Organizing Committee, Osaka, Japan, pp 3370–3380, https://aclanthology.org/C16-1318

de Vries H, Strub F, Chandar S, Pietquin O, Larochelle H, Courville AC (2017) Guesswhat?! visual object discovery through multi-modal dialogue. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, pp 4466–4475, https://doi.org/10.1109/CVPR.2017.475

Walker MA, Litman DJ, Kamm CA, Abella A (1997) PARADISE: A framework for evaluating spoken dialogue agents. In: 35th annual meeting of the association for computational linguistics and 8th conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Madrid, Spain, pp 271–280, https://doi.org/10.3115/976909.979652

Wan M, McAuley J (2016) Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In: 2016 IEEE 16th international conference on data mining (ICDM), IEEE, pp 489–498

Wang H, Peng B, Wong KF (2020a) Learning efficient dialogue policy from demonstrations through shaping. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 6355–6365, https://doi.org/10.18653/v1/2020.acl-main.566

Wang K, Tian J, Wang R, Quan X, Yu J (2020b) Multi-domain dialogue acts and response co-generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 7125–7134, https://doi.org/10.18653/v1/2020.acl-main.638

Wang L, Li J, Zeng X, Zhang H, Wong KF (2020c) Continuity of topic, interaction, and query: Learning to quote in online conversations. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 6640–6650, https://doi.org/10.18653/v1/2020.emnlp-main.538

Wang S, Zhou K, Lai K, Shen J (2020d) Task-completion dialogue policy learning via Monte Carlo tree search with dueling network. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3461–3471, https://doi.org/10.18653/v1/2020.emnlp-main.278

Wang W, Zhang J, Li Q, Hwang MY, Zong C, Li Z (2019a) Incremental learning from scratch for task-oriented dialogue systems. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3710–3720, https://doi.org/10.18653/v1/P19-1361

Wang X, Yuan C (2016) Recent advances on human-computer dialogue. CAAI Trans Intell Technol 1(4):303–312

Wang X, Shi W, Kim R, Oh Y, Yang S, Zhang J, Yu Z (2019b) Persuasion for good: Towards a personalized persuasive dialogue system for social good. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5635–5649, https://doi.org/10.18653/v1/P19-1566

Wang Y, Shen Y, Jin H (2018) A bi-model based RNN semantic frame parsing model for intent detection and slot filling. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 2 (short papers), association for computational linguistics, New Orleans, Louisiana, pp 309–314, https://doi.org/10.18653/v1/N18-2050

Wang Y, Guo Y, Zhu S (2020e) Slot attention with value normalization for multi-domain dialogue state tracking. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3019–3028, https://doi.org/10.18653/v1/2020.emnlp-main.243

Wang Y, Joty S, Lyu M, King I, Xiong C, Hoi SC (2020f) VD-BERT: A Unified Vision and Dialog Transformer with BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 3325–3338, https://doi.org/10.18653/v1/2020.emnlp-main.269

Wang Z, Zhang J, Feng J, Chen Z (2014) Knowledge graph embedding by translating on hyperplanes. In: Brodley CE, Stone P (eds) Proceedings of the twenty-eighth AAAI conference on artificial intelligence, July 27–31, 2014, Québec City, Québec, Canada, AAAI Press, pp 1112–1119, http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531

Wang Z, Schaul T, Hessel M, van Hasselt H, Lanctot M, de Freitas N (2016) Dueling network architectures for deep reinforcement learning. In: Balcan M, Weinberger KQ (eds) Proceedings of the 33nd international conference on machine learning, ICML 2016, New York City, NY, USA, June 19–24, 2016, JMLR.org, JMLR Workshop and Conference Proceedings, vol 48, pp 1995–2003, http://proceedings.mlr.press/v48/wangf16.html

Wang Z, Ho S, Cambria E (2020) A review of emotion sensing: Categorization models and algorithms. Multimedia Tools Appl 79:35553–35582

Welleck S, Weston J, Szlam A, Cho K (2019) Dialogue natural language inference. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 3731–3741, https://doi.org/10.18653/v1/P19-1363

Wen TH, Gašić M, Kim D, Mrkšić N, Su PH, Vandyke D, Young S (2015a) Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In: Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue, association for computational linguistics, Prague, Czech Republic, pp 275–284, https://doi.org/10.18653/v1/W15-4639

Wen TH, Gašić M, Mrkšić N, Su PH, Vandyke D, Young S (2015b) Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In: Proceedings of the 2015 conference on empirical methods in natural language processing, association for computational linguistics, Lisbon, Portugal, pp 1711–1721, https://doi.org/10.18653/v1/D15-1199

Wen TH, Gašić M, Mrkšić N, Rojas-Barahona LM, Su PH, Ultes S, Vandyke D, Young S (2016a) Conditional generation and snapshot learning in neural dialogue systems. In: Proceedings of the 2016 conference on empirical methods in natural language processing, association for computational linguistics, Austin, Texas, pp 2153–2162, https://doi.org/10.18653/v1/D16-1233

Wen TH, Gašić M, Mrkšić N, Rojas-Barahona LM, Su PH, Vandyke D, Young S (2016b) Multi-domain neural network language generation for spoken dialogue systems. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, association for computational linguistics, San Diego, California, pp 120–129, https://doi.org/10.18653/v1/N16-1015

Wen TH, Vandyke D, Mrkšić N, Gašić M, Rojas-Barahona LM, Su PH, Ultes S, Young S (2017) A network-based end-to-end trainable task-oriented dialogue system. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 1, long papers, association for computational linguistics, Valencia, Spain, pp 438–449, https://aclanthology.org/E17-1042

Weston J, Chopra S, Bordes A (2015) Memory networks. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings

Williams J (2013) Multi-domain learning and generalization in dialog state tracking. In: Proceedings of the SIGDIAL 2013 conference, association for computational linguistics, Metz, France, pp 433–441, https://aclanthology.org/W13-4068

Williams J, Raux A, Ramachandran D, Black A (2013) The dialog state tracking challenge. In: Proceedings of the SIGDIAL 2013 conference, association for computational linguistics, Metz, France, pp 404–413, https://aclanthology.org/W13-4065

Williams JD (2007) Partially observable markov decision processes for spoken dialogue management. PhD thesis, University of Cambridge

Williams JD (2014) Web-style ranking and SLU combination for dialog state tracking. In: Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL), association for computational linguistics, Philadelphia, PA, U.S.A., pp 282–291, https://doi.org/10.3115/v1/W14-4339

Williams JD, Zweig G (2016) End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. https://arxiv.org/abs/1606.01269

Williams JD, Asadi K, Zweig G (2017) Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Vancouver, Canada, pp 665–677, https://doi.org/10.18653/v1/P17-1062

Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8(3–4):229–256

Williams RJ, Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1(2):270–280

Wiseman S, Shieber S, Rush A (2017) Challenges in data-to-document generation. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, Copenhagen, Denmark, pp 2253–2263, https://doi.org/10.18653/v1/D17-1239

Wu CS, Xiong C (2020) Probing task-oriented dialogue representation from language models. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 5036–5051, https://doi.org/10.18653/v1/2020.emnlp-main.409

Wu CS, Madotto A, Hosseini-Asl E, Xiong C, Socher R, Fung P (2019a) Transferable multi-domain state generator for task-oriented dialogue systems. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 808–819, https://doi.org/10.18653/v1/P19-1078

Wu CS, Hoi S, Socher R, Xiong C (2020a) Tod-bert: Pre-trained natural language understanding for task-oriented dialogues. abs/2004.06871, https://arxiv.org/abs/2004.06871

Wu J, Wang X, Wang WY (2019b) Self-supervised dialogue learning. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3857–3867, https://doi.org/10.18653/v1/P19-1375

Wu S, Li Y, Zhang D, Zhou Y, Wu Z (2020b) Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 5811–5820, https://doi.org/10.18653/v1/2020.acl-main.515

Wu W, Guo Z, Zhou X, Wu H, Zhang X, Lian R, Wang H (2019c) Proactive human-machine conversation with explicit conversation goals. https://arxiv.org/abs/1906.05572

Wu Y, Wu W, Xing C, Zhou M, Li Z (2017) Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, pp 496–505, https://doi.org/10.18653/v1/P17-1046

Wu Z, Galley M, Brockett C, Zhang Y, Gao X, Quirk C, Koncel-Kedziorski R, Gao J, Hajishirzi H, Ostendorf M, et al. (2020c) A controllable model of grounded response generation. https://arxiv.org/abs/2005.00613

Xiao H, Huang M, Hao Y, Zhu X (2015) Transg: A generative mixture model for knowledge graph embedding. abs/1509.05488, https://arxiv.org/abs/1509.05488

Xiao H, Huang M, Meng L, Zhu X (2017) SSP: semantic space projection for knowledge graph embedding with text descriptions. In: Singh SP, Markovitch S (eds) Proceedings of the thirty-first AAAI conference on artificial intelligence, February 4–9, 2017, San Francisco, California, USA, AAAI Press, pp 3104–3110, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14306

Xie R, Liu Z, Jia J, Luan H, Sun M (2016) Representation learning of knowledge graphs with entity descriptions. In: Schuurmans D, Wellman MP (eds) Proceedings of the Thirtieth AAAI conference on artificial intelligence, February 12–17, 2016, Phoenix, Arizona, USA, AAAI Press, pp 2659–2665, http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12216

Xing C, Wu W, Wu Y, Liu J, Huang Y, Zhou M, Ma W (2017) Topic aware neural response generation. In: Singh SP, Markovitch S (eds) Proceedings of the Thirty-First AAAI conference on artificial intelligence, February 4–9, 2017, San Francisco, California, USA, AAAI Press, pp 3351–3357, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14563

Xing C, Wu Y, Wu W, Huang Y, Zhou M (2018) Hierarchical recurrent attention network for response generation. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, pp 5610–5617, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16510

Xu C, Wu W, Tao C, Hu H, Schuerman M, Wang Y (2019) Neural response generation with meta-words. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5416–5426, https://doi.org/10.18653/v1/P19-1538

Xu J, Wang H, Niu ZY, Wu H, Che W, Liu T (2020a) Conversational graph grounded policy learning for open-domain conversation generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 1835–1845, https://doi.org/10.18653/v1/2020.acl-main.166

Xu K, Tan H, Song L, Wu H, Zhang H, Song L, Yu D (2020b) Semantic Role Labeling Guided Multi-turn Dialogue ReWriter. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 6632–6639, https://doi.org/10.18653/v1/2020.emnlp-main.537

Yadollahi A, Shahraki AG, Zaiane OR (2017) Current state of text sentiment analysis from opinion to emotion mining. ACM Comput Surv (CSUR) 50(2):1–33

Yang S, Zhang R, Erfani S (2020) 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), Association for Computational Linguistics, Online, pp 1878–1888, https://doi.org/10.18653/v1/2020.emnlp-main.147

Yann D, Tur G, Hakkani-Tur D, Heck L (2014) Zero-shot learning and clustering for semantic utterance classification using deep learning. In: International conference on learning representations (cited on page 28)

Yao K, Zweig G, Hwang MY, Shi Y, Yu D (2013) Recurrent neural networks for language understanding. In: Interspeech, pp 2524–2528

Yao K, Peng B, Zhang Y, Yu D, Zweig G, Shi Y (2014) Spoken language understanding using long short-term memory neural networks. In: 2014 IEEE spoken language technology workshop (SLT), IEEE, pp 189–194

Yao K, Peng B, Zweig G, Wong KF (2016) An attentional neural conversation model with improved specificity. urlhttps://arxiv.org/abs/1606.01292

Yih Wt, He X, Gao J (2015) Deep learning and continuous representations for natural language processing. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts, Association for Computational Linguistics, Denver, Colorado, pp 6–8, https://doi.org/10.3115/v1/N15-4004

Yin J, Jiang X, Lu Z, Shang L, Li H, Li X (2016) Neural generative question answering. In: Kambhampati S (ed) Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI

2016, New York, NY, USA, 9-15 July 2016, IJCAI/AAAI Press, pp 2972–2978, http://www.ijcai.org/Abstract/16/422

Yoshino K, Hori C, Perez J, D'Haro LF, Polymenakos L, Gunasekara C, Lasecki WS, Kummerfeld J, Galley M, Brockett C, et al. (2018) The 7th dialog system technology challenge

Young S, Gašić M, Keizer S, Mairesse F, Schatzmann J, Thomson B, Yu K (2010) The hidden information state model: a practical framework for pomdp-based spoken dialogue management. Comput Speech Lang 24(2):150–174

Young T, Cambria E, Chaturvedi I, Zhou H, Biswas S, Huang M (2018) Augmenting end-to-end dialogue systems with commonsense knowledge. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, pp 4970–4977, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16573

Young T, Pandelea V, Poria S, Cambria E (2020) Dialogue systems with audio context. Neurocomputing 388:102–109

Young T, Xing F, Pandelea V, Ni J, Cambria E (2022) Fusing task-oriented and open-domain dialogues in conversational agents. Proceedings of the AAAI Conference on Artificial Intelligence 36:11622–11629

Yu F, Tang J, Yin W, Sun Y, Tian H, Wu H, Wang H (2020) Ernie-vil: Knowledge enhanced vision-language representations through scene graph. https://arxiv.org/abs/2006.16934

Yu T, Joty S (2020) Online conversation disentanglement with pointer networks. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 6321–6330, https://doi.org/10.18653/v1/2020.emnlp-main.512

Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontañón S, Pham P, Ravula A, Wang Q, Yang L, Ahmed A (2020) Big bird: Transformers for longer sequences. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual, https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html

Zahiri SM, Choi JD (2017) Emotion detection on tv show transcripts with sequence-based convolutional neural networks. https://arxiv.org/abs/1708.04299

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision, Springer, pp 818–833

Zhang C, Li Y, Du N, Fan W, Yu P (2019a) Joint slot filling and intent detection via capsule neural networks. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5259–5267, https://doi.org/10.18653/v1/P19-1519

Zhang C, Li Y, Du N, Fan W, Yu P (2019b) Joint slot filling and intent detection via capsule neural networks. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 5259–5267, https://doi.org/10.18653/v1/P19-1519

Zhang H, Lan Y, Pang L, Guo J, Cheng X (2019c) ReCoSa: detecting the relevant contexts with self-attention for multi-turn dialogue generation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3721–3730, https://doi.org/10.18653/v1/P19-1362

Zhang H, Liu Z, Xiong C, Liu Z (2020a) Grounded conversation generation as guided traverses in commonsense knowledge graphs. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 2031–2043, https://doi.org/10.18653/v1/2020.acl-main.184

Zhang J, Danescu-Niculescu-Mizil C (2020) Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 5276–5289, https://doi.org/10.18653/v1/2020.acl-main.470

Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J (2018a) Personalizing dialogue agents: I have a dog, do you have pets too? In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Melbourne, Australia, pp 2204–2213, https://doi.org/10.18653/v1/P18-1205

Zhang Y, Wallace B (2017) A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In: Proceedings of the eighth international joint conference on natural language processing (volume 1: long papers), Asian Federation of natural language processing, Taipei, Taiwan, pp 253–263, https://aclanthology.org/I17-1026

Zhang Y, Galley M, Gao J, Gan Z, Li X, Brockett C, Dolan B (2018b) Generating informative and diverse conversational responses via adversarial information maximization. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31: annual conference on neural information processing systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp 1815–1825, https://proceedings.neurips.cc/paper/2018/hash/23ce1851341ec1fa9e0c259de10bf87c-Abstract.html

Zhang Y, Ou Z, Hu M, Feng J (2020b) A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 9207–9219, https://doi.org/10.18653/v1/2020.emnlp-main.740

Zhang Z, Li J, Zhu P, Zhao H, Liu G (2018c) Modeling multi-turn conversation with deep utterance aggregation. In: Proceedings of the 27th international conference on computational linguistics, association for computational linguistics, Santa Fe, New Mexico, USA, pp 3740–3752, https://aclanthology.org/C18-1317

Zhang Z, Li X, Gao J, Chen E (2019d) Budgeted policy learning for task-oriented dialogue systems. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3742–3751, https://doi.org/10.18653/v1/P19-1364

Zhao T, Eskenazi M (2016) Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In: Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue, association for computational linguistics, Los Angeles, pp 1–10, https://doi.org/10.18653/v1/W16-3601

Zhao T, Eskenazi M (2018) Zero-shot dialog generation with cross-domain latent actions. In: Proceedings of the 19th annual sigdial meeting on discourse and dialogue, association for computational linguistics, Melbourne, Australia, pp 1–10, https://doi.org/10.18653/v1/W18-5001

Zhao T, Lee K, Eskenazi M (2018) Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), association for computational linguistics, Melbourne, Australia, pp 1098–1107, https://doi.org/10.18653/v1/P18-1101

Zhao T, Lala D, Kawahara T (2020a) Designing precise and robust dialogue response evaluators. In: Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics, online, pp 26–33, https://doi.org/10.18653/v1/2020.acl-main.4

Zhao X, Wu W, Xu C, Tao C, Zhao D, Yan R (2020b) Knowledge-grounded dialogue generation with pre-trained language models. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3377–3390, https://doi.org/10.18653/v1/2020.emnlp-main.272

Zhong P, Zhang C, Wang H, Liu Y, Miao C (2020) Towards persona-based empathetic conversational models. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 6556–6566, https://doi.org/10.18653/v1/2020.emnlp-main.531

Zhou H, Huang M, Zhu X (2016) Context-aware natural language generation for spoken dialogue systems. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, The COLING 2016 Organizing Committee, Osaka, Japan, pp 2032–2041, https://aclanthology.org/C16-1191

Zhou H, Zheng C, Huang K, Huang M, Zhu X (2020a) 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, association for computational linguistics, online, pp 7098–7108, https://doi.org/10.18653/v1/2020.acl-main.635

Zhou K, Prabhumoye S, Black AW (2018) A dataset for document grounded conversations. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics, Brussels, Belgium, pp 708–713, https://doi.org/10.18653/v1/D18-1076

Zhou L, Palangi H, Zhang L, Hu H, Corso JJ, Gao J (2020b) Unified vision-language pre-training for image captioning and VQA. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, AAAI Press, pp 13041–13049, https://aaai.org/ojs/index.php/AAAI/article/view/7005

Zhou X, Wang WY (2018) MojiTalk: Generating emotional responses at scale. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), association

for computational linguistics, Melbourne, Australia, pp 1128–1137, https://doi.org/10.18653/v1/P18-1104

Zhu Q, Cui L, Zhang WN, Wei F, Liu T (2019) Retrieval-enhanced adversarial training for neural response generation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics, Florence, Italy, pp 3763–3773, https://doi.org/10.18653/v1/P19-1366

Zhu Q, Zhang WN, Liu T, Wang WY (2020) Counterfactual off-policy training for neural dialogue generation. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 3438–3448, https://doi.org/10.18653/v1/2020.emnlp-main.276

Authors and Affiliations

Jinjie Ni[1] · Tom Young[1] · Vlad Pandelea[1] · Fuzhao Xue[1] · Erik Cambria[1]

Jinjie Ni
jinjie001@e.ntu.edu.sg

Tom Young
yang0552@e.ntu.edu.sg

Vlad Pandelea
vlad.pandelea@ntu.edu.sg

Fuzhao Xue
fuzhao001@e.ntu.edu.sg

[1] Nanyang Technological University, Singapore, Singapore