AI Open 3 (2022) 14–28

Contents lists available at ScienceDirect

AI Open

journal homepage: www.sciencedirect.com/journal/ai-open

# Learning towards conversational AI: A survey

Tingchen Fu [a], Shen Gao [b], Xueliang Zhao [b], Ji-rong Wen [a], Rui Yan [a,*]

[a] Gaoling School of Artificial Intelligence, Renmin University of China, China
[b] Wangxuan Institute of Computer Technology, Peking University, Beijing, China

ARTICLE INFO

*Keywords:*
Human-machine conversation
Response generation
Informativeness dialogue
Controllable dialogue

ABSTRACT

Recent years have witnessed a surge of interest in the field of open-domain dialogue. Thanks to the rapid development of social media, large dialogue corpus from the Internet builds up a fundamental premise for data-driven dialogue model. The breakthrough in neural network also brings new ideas to researchers in AI and NLP. A great number of new techniques and methods therefore came into being. In this paper, we review some of the most representative works in recent years and divide existing prevailing frameworks for a dialogue model into three categories. We further analyze the trend of development for open-domain dialogue and summarize the goal of an open-domain dialogue system in two aspects, informative and controllable. The methods we review in this paper are selected according to our unique perspectives and by no means complete. Rather, we hope this servery could benefit NLP community for future research in open-domain dialogue.

## 1. Introduction

Originated from the ELIZA (Shum et al., 2018) in the last century, open-domain conversation is a fascinating and exciting topic that continues to pique the curiosity of scholars. Researchers' dedication has resulted in a significant development, with a huge number of successful research works being translated and used in industry, resulting in concrete goods that help us in our everyday lives. One of such famous products is the Xiaoice[1] by Microsoft. Released in 2014, the intelligent chatbot has more than 660 million users around the world.

In fact, the research and application of dialogue in natural language date back further earlier to logical dialogue games and task-oriented dialogue. In contrast with computational approaches nowadays, a dialogue game defines a finite set of rules restricting the permissible dialogue act (Yuan et al., 2011; Lewin and Lane, 2000; McBurney and Parsons, 2003). Similarly, task-oriented dialogue usually formulates dialogue as a slot-filling task, with predefined user intent and main content. A task-oriented dialogue system is usually a scenario-specific system, assisting users to finish one or two kinds of specific tasks such as hotel booking or restaurant recommendation, useful in a vertical area and the response is confined in limited hypothesis space. Compared with its task-oriented counterpart, open-domain dialogue system has no explicit limitation on conversation topics and therefore has a much larger hypothesis space. Rather than assisting individuals in completing a specific task, the goal of an open-domain dialogue system seems implicit, that is, satisfying the human need for communication and association, acting as a digital companion to provide qualified and engaged conversation.

Though it has been explored for more than half a century, the problem of open-domain dialogue remains unsolved and challenging. And there is no fixed paradigm or standard procedure about how to establish a dialogue system. Hence we review some of the representative works and draw three kinds of most common frameworks. Namely, (1) retrieval-based method searches from a candidate pool and selects a proper candidate; (2) Generation-based method generates a response word by word from scratch or with a prompt; (3) Relatively newly-appeared hybrid method is more flexible and combines the strengths of both methods.

Besides, there is no golden rule or standard about what a good conversational system should be like. In this review, we posit that a human-like conversation system should be both (1) *informative* and (2) *controllable*. To be specific, a chatbot has to synthesize informative and meaningful responses to users. Always replying with safe answers like "I'm not sure about that" or "I don't know" gets users bored. To achieve

*Abbreviations:* PLM, pre-trained language model; KGC, knowledge-grounded conversation; LSTM, long-shot term memory; CNN, convolutional neural network; GRU, gated recurrent unit.

* Corresponding author.
*E-mail address:* lucas.futingchen@gmail.com (T. Fu).

[1] https://en.wikipedia.org/wiki/Xiaoice.

https://doi.org/10.1016/j.aiopen.2022.02.001
Received 19 October 2021; Received in revised form 17 February 2022; Accepted 22 February 2022
Available online 26 February 2022

this, we have to ground the dialogue system with external knowledge, in various formats. The act strategy and expression style of a dialogue system must be controllable so as to generate in the desired style or follow a specified strategy. It is also crucial to avoid generating offensive words or toxic phrases, which is an important step towards human-machine conversation. These two properties will be elaborated in the following sections, together with the efforts to achieve this.

There exist several surveys for the open-domain dialogue (Yan, 2018; Huang et al., 2020; Sun and Li, 2021; Tao et al., 2021). Similarly, this paper is a brief introduction and summary for representative related works and by no means complete, hoping to convenient future research work in this area.

The remainder of this paper is organized as follows (See Fig. 1): in Section 2, we formulate the task of open-domain dialogue and introduce the terminology in this field. In Section 3, we describe three most common frameworks for open-domain dialogue, namely retrieval-based, generation-based and hybrid methods. In Section 4 and Section 5, we summarize two goals for open-domain dialogue, and present some representative methods to accomplish these goals from different perspectives. In Section 6, we introduce several metrics to evaluate the performance of an open-domain dialogue. After that, we enumerate several corpus and benchmarks in Section 7 and finally in Section 8 we speculate the trend and direction for future research in this area.

## 2. Formulation and terminology

In exploration to the open-domain dialogue, researchers have developed a well-defined problem formulation. In general, we use the term *session* to describe a complete dialogue case. If a dialogue session only contains a single turn between two speakers, it is a *single-turn dialogue*, composed of a *query* and a *response* (Shang et al., 2015). Otherwise, if a dialogue session contains more than one turn, it is referred to as a *multi-turn dialogue* (Daniel and James, 2000). The last utterance is *response*, with the other utterances as the *context*. If there is only two interlocutors in a dialogue, it is a *dyadic dialogue*. Otherwise it is referred to as a *multi-party dialogue* (Hsueh et al., 2006).

Apart from conversational data, there may be also other external resources in various formats. In knowledge grounded conversation, the external resources are structured knowledge documents, knowledge graph, or unstructured knowledge sentences from Wikipedia. While in the multi-modality setting, it could be an image, a video clip, or a blend of image and video, depending on the task definition. Also, it could be the profile of the speaker. Further discussion about external resources will be presented in the later section.

Given the conversational data and the external resources, the essence of the open-domain dialogue is to feedback response to a query from users. To be specific, given a dialogue context $C = X_1Y_1X_2Y_2\cdots X_{t-1}Y_{t-1}$ and the last utterance from one user or the query $X_t$ as input, a model should generate a $Y_t$ as a response to the user, with the help of external resources $S$.

Formally, an open-dialogue system could be formulated as:

$$\widehat{Y} = \underset{Y\in\Omega}{argmax}\mathcal{P}(Y|C,S), \tag{1}$$

where $\Omega$ is the search space for the question, and $S$ is the external source to ground the dialogue. For a retrieve-based dialogue system, the search space is a set of candidate responses $Y_1, Y_2, Y_3, \ldots, Y_n$, where $n$ is usually equal or proportional to the size of dataset. For a generation-based dialogue system, the search space is exponential to the size of vocabulary $|\mathcal{V}|$ and the legal length interval of the response $[l_{min}, l_{max}]$. The core of a model is a search and scoring function $\mathcal{P}$, which explores and searches in the hypothesis space for the best response $\widehat{Y}$. Since the search space is not the same for different models and approaches, we will summary the mainstream frameworks for generation-based method, retrieval-based method, and hybrid method in the next section.

## 3. Framework

Basically, there are three kinds of frameworks or paradigms for a open-domain dialogue system. The summarization and category of existing open-domain dialogue system is fundamentally based on the previous work (Yan, 2018; Huang et al., 2020), yet we provide new insights and update with recently published approaches as well. In this section, we will elaborate them respectively for our readers.

### 3.1. Framework I: retrieval-based method

Retrieval-based methods assume that the next dialogue utterance is mixed in a large set of candidate responses. Typically, all possible responses in the training set are used as the candidate set. A retrieval-based model must evaluate all of the replies in the candidate set and assign each one a score depending on whether or not it is the appropriate response for the current context. Finally, the candidate with the highest score is outputted as the response.

In a nutshell, the core of a retrieval-based model is a score function $s$ () and an encoding function $e$(). The encoding function is responsible for encoding the context and response from natural language into dense representations. And the role of the score function is to give a matching score for a pair of (context, response). Based on the form of encoding function and score function, existing retrieval-based methods could be divided into two categories: (1) shallow interaction and (2) deep interaction (Huang et al., 2020).

#### 3.1.1. Shallow interaction

Shallow interaction means that a candidate response and a query are encoded independently, agnostic to each other. The interaction between a query and a response begins after the encoding phase, so this paradigm is also named as representation-based methods by some surveys (Huang et al., 2020). The matching score for a pair of (query, response) is calculated based on the encoding of context and response, so the score function is a bivariate function and the retrieval model could be formulated as:

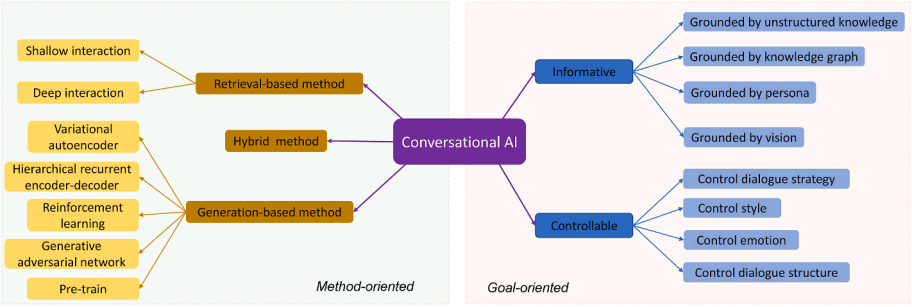


**Fig. 1.** Key issues and key techniques in open-domain dialogue systems.

$$\widehat{Y} = \underset{Y \in \mathcal{Y}}{argmaxs}(e(C), e(Y)), \tag{2}$$

where $\mathcal{Y}$ is the candidate set or the searching space. Early retrieval-based methods mostly fall in this category (Kang et al., 2014; Wang et al., 2015, 2017), tending to resort to lexical co-occurrence or syntax analysis. TF-IDF (Kang et al., 2014), the number of common words (Wang et al., 2017) and the dependency tree (Wang et al., 2015) are all widely used as score functions. In recent years, with a resurgent of neural network and deep learning, neural methods are gradually occupying the mainstream. Score functions in neural methods are usually bilinear function (Lu and Li, 2013), multi-layer perceptron (Hu et al., 2014) or Euclidean distance (Yang et al., 2018a). And the encoding function could be a convolutional neural network (CNN), a recurrent neural network (RNN), or a combination of both.

CNN is widely used to model sentences (Hu et al., 2014; Rakhlin, 2016; Shin et al., 2018) for convolutional neural network is good at extracting robust and abstract features of input. Hu et al. (2014) adopt CNN as encoding function and propose a new convolutional architecture for modeling sentences with a stack of convolution layers and max-pooling layers. To deal with variable-length sentences and eliminate boundary effect, a gate is added to the convolution layer that sets all the output vectors to zero if the input is all zero, which is an all-zero padding in nature.

Apart from CNN, RNN is expert at modeling time-dependent sequence like dialogue text. Lowe et al. (2015) exploit RNN to represent context and response with the last hidden state in RNN. The final matching score is calculated as the bilinear transformation of the hidden states (Bahdanau et al., 2015). Sometimes a single RNN is not enough maybe. In contrast with Lowe et al. (2015), enhanced sequential inference model (ESIM) (Chen and Wang, 2019), another shallow interaction method, devises a much more complicated and exquisite encoding function, which is helpful to model the interaction between utterance history and the response. The score function consists of a multi-layer perceptron (MLP) and a bidirectional long short-term memory (LSTM) network. Every utterance in a multi-turn dialogue session and each responses in candidate set are represented with a semantic vector and a dual vector to embody local matching information. The impressive results on Ubuntu Dialogue Corpus (Lowe et al., 2015) imply the success of this method.

However, Zhou et al. (2016) point out that neither CNN nor RNN is enough. RNN pays more attention to the word-level encoding, but ignores the interaction at the utterance level. Therefore it proposes a multi-view method, using a LSTM and a CNN to obtain the word-level encoding and the utterance-level encoding, respectively. Following the precedent of (Zhou et al., 2016), Yan et al. (2016), Tao et al. (2019), Li et al. (2021a) utilize both CNN and RNN for encoding.

#### 3.1.2. Deep interaction

We classify a retrieval-based method as deep interaction if the interaction of context and response begins during the encoding phase and the encoding representation incorporates mutual knowledge about each other. For deep interaction, the score function slightly differs from its shallow-interaction counterpart:

$$\widehat{Y} = \underset{Y \in \mathcal{Y}}{argmaxs}(e(C, Y)). \tag{3}$$

Note that the major difference lies in the form of encoding function $e(\cdot, \cdot)$, which indicates the fact that the context and the response interact with each other when encoding. A large proportion of methods in this field formulate a multi-turn dialogue as a sequence and every utterance in dialogue as an element (Wu et al., 2017; Zhou et al., 2018c; Yang et al., 2018a). Then the matching between the context and a response could be decomposed into the matching between each utterance and response. A representative method that has to be mentioned is sequential matching network (SMN) (Wu et al., 2017). The model first separately encodes each utterance in context and the response together using a shared-parameter RNN to obtain a matrix consisting of the hidden states of each time step. Another feature matrix is calculated by the similarity of words in context and response. Then a CNN is applied to convert several similarity matrices into a feature vector before another RNN finally converts the feature vector into a match score. There are a number of works following SMN(Zhou et al., 2018c; Yang et al., 2018a; Zhang et al., 2018d). The deep attention matching (Zhou et al., 2018c) is one of them. To make up for the shortcoming of RNN in capturing long-term and multi-grained semantic representation, the model replaces the first RNN in SMN with a five-layer hierarchically stacked self-attention. Deep matching network (DMN) (Yang et al., 2018a) is also an extension of SMN with external knowledge. It creatively introduces pseudo-relevance feedback and extends the candidate response with relevant words extracted from external knowledge. Deep utterance aggregation system (DUA) (Zhang et al., 2018d) extends SMN with a weighted context utterance. Specifically, the last utterance is the most relevant one and is concatenated with some previous utterances as well as the response to get a final representation.

Sequential matching methods view multiple utterances as a flat sequence, unable to discover the hierarchical structure or deep underlying pattern in the utterance history. To cope with this, similar to Zhou et al. (2018c), Tao et al. (2019) explore the representation of context-response pairs at different levels of granularity. It considers three levels of representation, namely the word representation, contextual representation and attention-based representation, obtained through CNN and RNN. A MLP is finally used as a score function. The hierarchical division in Tao et al. (2019) is purely based on lexical granularity, without considering the semantics embodied in n-grams or sub-sentences. Different from that, Xu et al. (2021) construct a topic-based hierarchical structure. It segments the utterance history into several groups according to the similarity of their encoding. After segments weighting and a stack of cross attention, the relation of different segments is modeled with a GRU following Wu et al. (2016). Though deep neural network with sophisticated structure achieves promising results (Tao et al., 2019), Zhang et al. (2021) challenge that deep neural network is hard to train because the gradient is difficult to propagate back to the bottom layer of the network. In light of this, they add a residual layer to allow the gradient to directly flow back.

All of the studies described above are concerned with how to fully utilize previous utterances in a dialogue session, assuming that all utterances are useful in predicting the future utterance. Yet Yuan et al. (2019) argue that less is more: detecting and removing noise in utterance history promote the performance of response retrieval. Specifically, it devises a multi-hop selector to model the utterance-level correlation between a context and a response. By selecting only relevant utterances in context, it generates a better representation of context and response for matching.

Recently, the great success of pre-trained language models (PLM) inspires new methods in retrieval-based dialogue. To utilized pre-trained language model, Henderson et al. (2019b) concatenate all the utterance history and the response together to form a long sequence and send it into BERT (Devlin et al., 2018). The representation for the token [CLS] is then utilized for predicting the matching score. Pre-trained language models like BERT are usually trained on general corpus like the BooksCorpus [? ], so there is usually a domain shift when directly fine-tuning them on the downstream tasks, which may hurt performance. In light of this, Whang et al. (2020) post-train BERT (Devlin et al., 2018) for response selection on Ubuntu Dialogue Corpus with next utterance prediction and masked language modeling as training objectives, borrowed from BERT. Similarly, Xu et al. (2020c) propose a set of self-supervised training objective to improve the coherence and consistency of the retrieved response.

Another shortcoming of BERT is, it is incompetent in detecting sentence order, since the training objectives of BERT do not put emphasis on this. To make up for this, Han et al. (2021) propose a new training

objective named utterance relevance classification, which adds the utterance history in a dialogue session to the response candidate set as negative samples. Whang et al. (2021) agree with Han et al. (2021) on the importance of learning sentence order. Instead of adding negative samples, it proposes novel utterance manipulation strategies including utterance insertion, utterance deletion and utterance search to help the BERT model to learn the underlying pattern of the sentence order.

In spite of the widely recognized capacity of the pre-trained language models, recently studies report that the contextualized words and sentence representations of PLM are anisotropic, occupying a narrow cone in the vector space. To ameliorate this, Li et al. (2021b) conduct fine-grained contrastive learning on both instance view and category view to extend the expressiveness of BERT. In addition, Zhao et al. (2019) and Qin et al. (2020) incorporate knowledge from external documents to guide the retrieval of the golden response.

To compare the performance of these methods, we collect their results on two widely used benchmarks, namely Ubuntu Corpus V1 (Lowe et al., 2015) and Douban Corpus (Wu et al., 2017). Hopefully, Table 1 could help our readers to observe the deviation of the scores and the efficacy of different approaches.

### 3.2. Framework II: generation-based method

Instead of selecting an answer from the candidate set, a generation-based method has to synthesize a response word by word. Existing frameworks of generation-based methods are mainly based on encoder-decoder architecture. Briefly, the encoder encodes the context into a hidden state, a vector representation of contextual information. The decoder then is responsible for picking a new word and updating the hidden state accordingly at each time step. Formally,

$$\widehat{Y} = \underset{Y\in\Omega}{argmax} \prod_{i=1}^{|l_Y|} p_\theta(y_i|C, y_{<i}), \tag{4}$$

where $p_\theta$ is a generation model parameterized by $\theta$ and $|l_Y|$ is the length of response. $y_i$ denotes the $i$-th token in the response and we refer to $y_{<i}$ as the tokens generated before the $i$-th timestep. This way of generation is also named auto-aggressive, as the decoder always refers to the previously generated words as a known condition when producing a new one. Dependency on the previous generated tokens makes it impossible to generate all the tokens in a parallel way, which is a bottleneck for speeding up generation process. Motivated by this, some researchers make efforts to study non-auto-aggressive generation paradigm (Kaiser et al., 2018; Lee et al., 2018), a new generation paradigm that relies on the hypothesis that every word is conditionally independent in per-step distribution.

$$\widehat{Y} = \underset{Y\in\Omega}{argmax} \prod_{i=1}^{|l_Y|} p_\theta(y_i|C) \tag{5}$$

Though non-auto-aggressive generation still has a gap in performance when compared with its auto-aggressive counterpart, it opens up a new research direction for generation-based methods.

The encoder-decoder framework is held true for both the recurrent neural network and transformer network, where the encoder and decoder are usually LSTM or GRU for the former and a stack of self-attention layers and cross-attention layers for the latter. In general, encoder-decoder framework is shared by many text generation tasks. As we can see, many techniques and learning methods in generation-based open-domain dialogue are borrowed from other areas (Sordoni et al., 2015a; Kingma and Welling, 2013).

#### 3.2.1. Hierarchical recurrent encoder-decoder

Hierarchical recurrent encoder-decoder (HRED) (Sordoni et al., 2015a) is designed for presenting query suggestions to the users of search engine at the very beginning. It uses two recurrent networks to represent query-level recurrent state and session-level recurrent state respectively. Intuitively, with the representation of utterance history at multiple levels, the model is keen on capturing both local and global semantic information.

Serban et al. (2016) are the first to transfer this idea from query suggestion to dialogue. The model includes two RNNs, an encoder RNN for mapping each utterance to an utterance vector and a higher-level context RNN that keeps track of the history utterance by processing each utterance vector in an iterative way.

However, the traditional HRED architecture directly using the last hidden state of higher-level RNN as the final context representation, abandoning all the hidden states of previous utterances. To better utilize all the hidden states, a number of researchers explore how to integrate all the higher-level hidden states (Yao et al., 2015; Serban et al., 2017; Sordoni et al., 2015b). Among them, Tian et al. (2017) propose WSeq, which weighted sums or weighted concatenates all the hidden states at the higher-level according to the similarity between the corresponding utterance history and the query. WSeq and other previous works strive to design a sophisticated encoder, yet HRAN (Xing et al., 2017) figures out that the decoder is also important. To be more specific, it weights each utterance hidden state by their similarity with the decoder hidden state to update the decoder hidden state at each time step. To combine the strengths of both WSeq and HRAN, Recosa (Zhang et al., 2019a) extends the traditional HRED with attention mechanism in both encoder side and decoder side. A positional encoding is additionally added to the utterance vector to help to recognize the order of utterance history.

**Table 1**
Retrieved-based models comparisons on two benchmarks, numbers in bold are the best results.

| Models | Ubuntu | | | Douban | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $R_1$@10 | $R_2$@10 | $R_5$@10 | MAP | MRR | P@1 | $R_1$@10 | $R_2$@10 | $R_5$@10 |
| TF-IDF (Lowe et al., 2015) | 0.410 | 0.545 | 0.708 | 0.331 | 0.359 | 0.180 | 0.096 | 0.172 | 0.405 |
| RNN (Lowe et al., 2015) | 0.403 | 0.547 | 0.819 | 0.390 | 0.422 | 0.208 | 0.118 | 0.223 | 0.589 |
| CNN (Kadlec et al., 2015) | 0.549 | 0.684 | 0.896 | 0.417 | 0.440 | 0.226 | 0.121 | 0.252 | 0.647 |
| LSTM (Kadlec et al., 2015) | 0.638 | 0.784 | 0.949 | 0.485 | 0.537 | 0.320 | 0.187 | 0.343 | 0.720 |
| SMN (Wu et al., 2017) | 0.726 | 0.847 | 0.961 | 0.529 | 0.569 | 0.397 | 0.233 | 0.396 | 0.724 |
| DUA (Zhang et al., 2018d) | 0.752 | 0.868 | 0.962 | 0.551 | 0.599 | 0.421 | 0.243 | 0.421 | 0.780 |
| DAM (Zhou et al., 2018c) | 0.767 | 0.874 | 0.969 | 0.550 | 0.601 | 0.427 | 0.254 | 0.410 | 0.757 |
| IoI (Tao et al., 2019) | 0.796 | 0.874 | 0.974 | 0.573 | 0.621 | 0.444 | 0.269 | 0.451 | 0.786 |
| MRFN (Tao et al., 2019) | 0.786 | 0.886 | 0.976 | 0.571 | 0.617 | 0.448 | 0.276 | 0.435 | 0.783 |
| ESIM (Chen and Wang, 2019) | 0.796 | 0.894 | 0.975 | – | – | – | – | – | – |
| MSN (Yuan et al., 2019) | 0.800 | 0.899 | 0.978 | 0.587 | 0.632 | 0.470 | 0.295 | 0.452 | 0.788 |
| BERT-VFT (Whang et al., 2020) | 0.855 | 0.928 | 0.985 | – | – | – | – | – | – |
| BERT-UMS (Whang et al., 2021) | 0.875 | 0.942 | 0.988 | 0.625 | 0.664 | 0.499 | 0.318 | 0.482 | 0.858 |
| BERT-SL (Xu et al., 2020c) | 0.884 | 0.946 | 0.990 | – | – | – | – | – | – |
| BERT-FC (Li et al., 2021b) | 0.886 | 0.948 | 0.990 | 0.627 | 0.670 | 0.500 | **0.326** | 0.512 | 0.869 |
| BERT-FP (Han et al., 2021) | **0.911** | **0.962** | **0.994** | **0.644** | **0.680** | **0.512** | 0.324 | **0.542** | **0.870** |

#### 3.2.2. Variational autoencoder

Variational autoencoder (VAE) (Kingma and Welling, 2013) is also a popular approach for dialogue generation, though it is proposed for image generation at the very beginning. It alters the autoencoder architecture by substituting the deterministic encoding function with a learned posterior recognition model, which parameterizes an estimated posterior distribution over the latent space. Intuitively, the VAE learns codes not as a single point, but as a soft ellipsoidal region in latent space, forcing the codes to fill the space rather than memorizing the training data one by one as isolated codes.

Bowman et al. (2015) are the first to successfully transfer VAE from computer vision to natural language processing and proposes a novel variational method. The final hidden state of the encoder is not directly sent for decoding but is used to determine the distribution of a latent variable. A sampling from the distribution is then sent to the decoder for synthesizing a response.

To enable the VAE to model long-term text, variational hierarchical recurrent encoder-decoder (VHRED) (Serban et al., 2017) combines the traditional HRED model and a latent variable at the context level. The latent variable subjects to a Gaussian distribution, whose normal mean and covariance matrix is calculated based on the last hidden state of the context RNN. Under the assumption of VHRED, no linguistic label is supplied and the model works in a purely unsupervised way. SPHRED (Shen et al., 2017) amends this assumption and studies the conditional generation where partial or all linguistic attribute labels are given, which are incorporated into the latent variable. When the attributes are unknown, a multi-layer perceptron or support vector machine is utilized to predict the attribute, whose precision is an additional training objective. Following the idea of SPHRED, Zhao et al. (2017) regards attributes label as extra knowledge. It proposes kgCVAE, incorporating linguistic features as extra knowledge for latent variables.

Besides, VAE brings uncertainty into the traditional encoder-decoder framework due to the stochastic sampling process of latent variables. Potentially, it is a prospective technique to boost the diversity of generation or model the one-to-many relationship in dialogue (Qiu et al., 2019; Gao et al., 2019a).

In spite of its broad prospect, the training process of latent variables is non-trivial. As pointed out in Bowman et al. (2015), VAE methods suffer from *vanishing latent variable* problem more or less. It means that the decoder only pays attention to the previous words, ignoring the latent variables, presumably due to the capacity of the decoder. To alleviate this, several tricks have been proposed:

- **KL-annealing** (Bowman et al., 2015): gradually increase the coefficient of the KL-divergence term from 0 to 1.
- **dropout word decoding** (Bowman et al., 2015): set a certain percentage of the target words to 0. But when the drop rate is too high, it may hurt the performance.
- **bag of word loss** (Zhao et al., 2017): require the decoder to correctly predict the bag of word in response.
- **mutual information maximization** (Zhao et al., 2018): augment the evidence lower bound objective with a mutual information term.
- **PI controller** (Shao et al., 2021): control the coefficient of the KL-divergence term according to the current KL-divergence.

#### 3.2.3. Reinforcement learning

Regular training objective like maximum likelihood estimation (MLE) tends to favor high-frequency response, which is a major factor of *safe response* problem (Li et al., 2016b). However, without MLE, directly optimizing the generation process is not easy. As the synthesis of every word is usually from sampling or beam search, the generation process is discrete. As a result, dialogue generation is undifferentiable in nature, hindering the backpropagation of the gradient.

To handle these barrier, reinforcement learning is a good choice. Besides, an additional advantage lies in the flexibility of self-designed reward, which could impose expected properties (i.e. fluency, coherence, and relevance) to the training of the dialogue system.

Nevertheless, reinforcement learning is notorious for slow convergence and high variance in the training process. To speed up convergence, warm-up is a practical method, with maximum likelihood estimation as a training objective. Since Q-learning directly estimates the future reward of each action, which is different from MLE in orders of magnitude, it is not appropriate in natural language generation. Therefore, policy gradient method (Sutton et al., 2000) is much more popular than Q-learning (Li et al., 2016b; Zhao et al., 2020b) when applying reinforcement learning in natural language generation.

Roughly speaking, the designed reward could be divided into two categories: learnable rewards and unlearnable rewards. Learnable rewards are usually the log-likelihood given by a language model or a discriminator, while the unlearnable rewards are probably lexical metrics like BLEU (Papineni et al., 2002) or F1 (Dinan et al., 2019).

For example, Li et al. (2016b) use the reward to encourage human-like response to be informative and coherent. The reward consists of three parts: ease of answer, information flow, and semantic coherence, all are learnable rewards given by a language model or the similarity of the encoder hidden state. The learnable reward in Li et al. (2016b) is given by a language model, whose parameters are frozen after warm-up. To allow the joint training in a unified framework, Zhang et al. (2018a) devise a dual learning framework featuring a generation model and a coherence model, and those two parts are optimized alternatively.

PRGDDA (Yang et al., 2018b) use both unlearnable reward and learnable reward to consider multiple factors in optimization. Its reward consists of four independent parts: reconstruction reward, the log-likelihood for the post generator to reconstruct a post from the response; language model reward, the log-likelihood for a pre-trained language model to generate the response; a topic coherence reward which is the cosine similarity between the topic vector presentation of both query and response; and finally a BLEU (Papineni et al., 2002) reward that measures the BLEU score between the generated hypothesis and the golden answer. Similar to PRGDDA, the reward of transmitter in $\mathcal{P}^2$ bot (Liu et al., 2020) is also composed of four parts: language style reward, the length-normalized log-likelihood given by a pre-trained language model (i.e. GPT); discourse coherence reward, the log probability of the generated response be recognized as the next utterance to the query given by a pre-trained predictor; mutual persona perception reward, a long-term reward measuring the mutual perception between the interlocutors.

Learnable rewards require training an additional reward model, which is crucial to find the optimal solution. Thus a number of methods resort to only unlearnable rewards. For instance, Gao et al. (2019a) calculated the reward as unigram F1 (Dinan et al., 2019) between the generated hypothesis and the golden answer. The reward connects the inference network and the generation network, overcoming the undifferentiable process of latent variables and allows the joint optimization of these two components. Unigram F1 reward is also adopted by KnowledGPT (Zhao et al., 2020b). KnowledGPT is composed of two components, a knowledge selector and a generator. After the selector chooses a piece of knowledge from a knowledge pool, the generator synthesizes a response based on the selected knowledge. The unigram F1 between the synthesized response and the golden response is used as the reward for the selector.

The above methods mostly exploit reinforcement learning for promoting coherence or context relevance. But the application of reinforcement learning is not limited to this. Saleh et al. (2020) explore to tackle the repetition and toxicity in dialogue generation. To be specific, it utilizes See et al. (2019) to detect repetition in dialogue, which is simple but works in a pragmatic way. Naive Bayes logistic regression classifier (Saleh et al., 2019) is exploited to recognize the toxic word. In architecture, the model is based on VHRED, and reward takes its effect to a higher level and permits the gradient flow from decoder RNN to encoder RNN, and finally both flow through the lower-level encoder

RNN.

#### 3.2.4. Generative Adversarial Network

One of the ultimate goals of a dialogue system is to produce a human-like conversation. Thus a good dialogue system should be able to generate a response that is indistinguishable from human conversation. A large group of researchers views Generative Adversarial Network (GAN) (Goodfellow et al., 2014) as a proper method to achieve this goal. Since the generation process is discrete in nature, the interaction between the generator and the discriminator is usually undifferentiable. To cope with this, reinforcement learning is then applied (Liu et al., 2020) and a specially designed reward acts as the signal between the generation model and the adversarial model.

Li et al. (2017a) is the first work to successfully introduce GAN into natural language processing and achieves promising results. The model architecture is straightforward: a generative model to generate a response given a query and a discrimination model to distinguish the matching or mismatching between a given query and response. It is notable that the authors decompose the reward for the whole generated hypothesis into *Reward for Every Generation Step* for more accurate reward at a finer granularity. Other than reinforcement learning, Xu et al. (2017) try to build the connection between the generation model and the discrimination model with another method. They develop an approximate embedding layer, directly converting the prediction of the next token into a mixed word embedding, whose mixing weight is the probability given by the softmax layer of the decoder. A straightforward design of discriminator is a classifier, outputting a confidence score as result. However, Xu et al. (2018) find that a classifier-based discriminator has a high probability to suffer from the saturation problem and fails to work. Thus, a language-model-based discriminator is proposed and cross-entropy is used as a reward. Adversarial Information Maximization (AIM) (Zhang et al., 2018c) also explores another kind of discriminator, whose training objective for its discriminator is to maximize the cosine distance of a pair of positive examples and at the same time minimize the cosine distance of a pair of negative examples. The cosine distance is measured in a shared embedding space. In DialogWAE (Gu et al., 2018), however, the loss of discriminator plays a similar role as the KL-divergence loss in VAE methods. It is a novel variant of VAE-based method, training a GAN within the latent space to distinguish the prior samples and the posterior samples. Recently, Feng et al. (2020) put forward a new framework with a forward discriminator and a backward discriminator. It is worth noting that this work introduces *future information* and reformulates the context-response pair to triplets in the format of (context, response, future). So the two discriminators are responsible for checking the forward pass (context → response) and the backward pass (future → response) respectively.

#### 3.2.5. Pre-trained language model

The great success of BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020a) in various tasks and benchmarks sparks the interest of NLP researchers and language model pre-training is gradually viewed as an effective method in many sub-fields in open-domain dialogue. Basically, pre-training and fine-tuning are the two main steps of these types of techniques. In the pre-training phase, a large-scale corpus is utilized to optimize all the parameters without any task-specific objectives. In fine-tune phase, part (or all) of the parameters are further optimized to fit in a downstream task.

DialoGPT (Zhang et al., 2019d) inherits GPT-2 in model architecture and is trained with comment chains scraped from Reddit. Concretely, they model a multi-turn dialogue session as a long text and frame the generation task as language modeling. Mutual information maximization is adopted as a re-ranking trick to avoid safe response. Though attaining surprising results in many benchmarks, previous methods usually model the dialogue as a one-to-one problem. Therefore, PLATO-2 (Bao et al., 2019) is proposed to tackle the inherent one-to-many problem in dialogue. To be specific, it models the one-to-many relationship via a $K$-way categorical latent variable to indicate particular speech act. Finally, an evaluation model is responsible to pick the best responses from many candidates. Meena (Adiwardana et al., 2020) further scales up the parameters in network to 2.6B and more social media conversations. As a result, it surpasses the DialoGPT significantly. Besides, it demonstrates that an end-to-end neural model with sufficiently low perplexity could achieve impressive results. The training corpus of pre-training model usually comes from Wikipedia, which is clean but tedious and lacking in expression style. Inspired by the Blended Skill Talk, Roller et al. (2020) try to make a difference. They argue that only a single corpus is not enough and introduces various corpus focusing on different aspects of a dialogue system. With diverse corpus, the model is able to learn various conversation skills including showing abundant knowledge, keeping a consistent persona and so on. Different from most of methods that train a pre-training model as a generator (Zhang et al., 2019d; Adiwardana et al., 2020; Roller et al., 2020), ELECTRA (Clark et al., 2020) pre-trains a text encoder as a discriminator. Specifically, it replaces some words in a sentence with plausible alternatives and requires the model to distinguish the replaced word. Compared with the original masked language modeling, this objective is more effective as it not only considers the masked positions but all the words in a sentence.

### 3.3. Framework III: hybrid method

Retrieval-based methods search from a pre-defined candidate set, thus the response tends to be of high quality, fluent and grammatical. Yet the hypothesis space is limited and crucial to the success of retrieval. Generation-based methods, on the other hand, enjoy a large search space and have the possibility to produce novel and unseen responses. The cost is there is no grantee for the quality of generated response. Recently, some works try to seek a new paradigm that combines the strengths of both frameworks. They usually take a two-stage procedure. In the first stage, a number of similar conversations instances are retrieved from the dataset. In the second stage, the retrieved instances are exploited to assist the generator in various ways. In all, hybrid methods could be formulated as:

$$\widehat{Y} = \underset{Y \in \Omega}{argmax} \prod_{i=1}^{|l_Y|} p_\theta(y_i|\mathcal{P}), \tag{6}$$

where $\mathcal{P}$ is a set of retrieved instances, serving as prototypes or skeletons.

The first hybrid method is proposed by Weston et al. (2018). They put up a straightforward method named RetrieveNRefine. It first retrieves a response with key memory network (Miller et al., 2016). Then the retrieved response and the context are concatenated together and sent into a two-layer LSTM for generation. Overall, Weston et al. (2018) is more like a generation method since the final response is always from generation though sometimes a retrieved one may be a better choice. Inspired by this, Song et al. (2018) re-ranked the retrieved responses and the generated one together and the best response is outputted as the final answer. Weston et al. (2018) and Song et al. (2018) directly send the retrieved response to the generator, without further consideration about the difference between the retrieved instances and the current query, which might play an important role in generation. In light of this, Pandey et al. (2018) retrieve similar (context, response) pairs from a corpus and calculate an exemplar vector for every (context, response) pair based on the retrieved responses and the current context. Then the exemplar vectors are summed up to assist the generation process. Following Pandey et al. (2018), Wu et al. (2019) propose a new paradigm named prototype-then-edit. Given a context, It first uses a selector to retrieve similar (context, response) pairs from the candidate set, which they name as prototype. An edit vector is computed according to the difference between the prototype contexts and the current context. Finally, a RNN generates a response based on the edit vector and the

prototype response. Though these methods achieve delightful results, Cai et al. (2018) find that in previous hybrid methods (Weston et al., 2014; Pandey et al., 2018), the decoder tends to direct repeat the retrieved prototype without modification, resulting in improper responses. They attribute this phenomenon to the useless noise in the retrieved prototype, which usually contains useless entities. Therefore they propose a skeleton-to-response framework, which inserts new words and deletes unrelated words in the prototype and only keeps a skeleton for the subsequent generation. The combination of retrieval and generation is also helpful to update the knowledge of the model. Lewis et al. (2020b) introduce retrieval-augmented generation (RAG) for knowledge-intensive dialogue. The model exploits both parametric memory from the parameter in pre-trained language model and non-parametric memory in form of a dense vector index of Wikipedia.

## 4. Informative: grounded by external knowledge

Open-domain dialogue models often suffer from the safe response problem (Zhang et al., 2018c), in other words, they usually generate bland or generic responses like "I'm not sure", "I don't know" or similar. It's hardly surprising, given that a conversation model's parameters may not always include the essential knowledge. In recent years, there is a tendency (Dinan et al., 2019; Kim et al., 2020; Meng et al., 2020) to supply external knowledge to ground the dialogue. Based on the form of external knowledge, there are generally 4 types of knowledge-grounded conversation (KGC): (1) Grounded by unstructured knowledge sentences and documents; (2) Grounded by structured knowledge graph; (3) Grounded by user profiles (persona); (4) Grounded by visual information.

### 4.1. Grounded by unstructured knowledge

The task of dialogue grounded by unstructured knowledge is to build a *wizard of Wikipedia* (Dinan et al., 2019). Given a dialogue context and necessary unstructured knowledge sentences, it should be able to first select a proper knowledge from a knowledge pool, and then generate a response based on the selected knowledge.

Ghazvininejad et al. (2018) explore this problem with a memory network (Weston et al., 2014). The knowledge sentences are represented with a bag-of-word vector in a fixed size, and the knowledge in memory is selected with attention mechanism (Bahdanau et al., 2015). To help the model select a proper knowledge, reconstruction of knowledge facts is used as a training objective. After Ghazvininejad et al. (2018), Dinan et al. (2019) substitute the memory network with a transformer (Vaswani et al., 2017). A highlight of this work is the introduction of the knowledge selection loss, whose role is guiding the model to select the golden knowledge. The knowledge selection loss is widely adopted by later works (Kim et al., 2020; Meng et al., 2021). However, vanilla transformer decoder is unable to model the attention on the context and knowledge respectively. ITDD (Li et al., 2019) revises the architecture of transformer to conduct knowledge attention and context attention at every encoder layer. Besides, it draws inspiration from deliberation network (Xia et al., 2017) and adopts a two-pass generation. The first pass pays attention to context, whereas the second pass is concerned with external knowledge.

The methods described above either implicitly select knowledge with attention mechanism (Ghazvininejad et al., 2018; Li et al., 2019) or view the knowledge selection as a deterministic process (Dinan et al., 2019). However, dialogue is one-to-many and uncertain in nature. Especially, The hypothesis space is largely extended when ample external knowledge is provided. When people choose different knowledge sentences, they usually give multiple diverse responses. Therefore some researchers view knowledge selection as a probabilistic process and resort to variational autoencoder (Kingma and Welling, 2013) since variational autoencoder is an effectual tool to model a stochastic process. Latent variables are introduced to indicate the golden knowledge. For example, Lian et al. (2019) propose Posterior Knowledge Selection model (PostKS) featured with a prior knowledge module and posterior knowledge module. The main motivation is that the response is a semantic reflection of the selected knowledge. Thus the clue in response is helpful for the model to find the corresponding knowledge. The gains from the posterior module are therefore distilled to improve the prior module. However, in a multi-turn dialogue session, the knowledge selection is dynamic with the conversation flow and the speaker may choose different knowledge at different turns. Therefore SKT (Kim et al., 2020) uses sequential latent variables to dynamically select knowledge at each turn of dialogue. The posterior network samples a knowledge sentence at every turn, and the representation of the sampled knowledge is further utilized to update the parameters in the posterior network and the prior network. For optimization, an auxiliary knowledge loss is added to the original evidence lower bound objective, which is calculated as the log-likelihood of choosing the golden knowledge. Meng et al. (2020) also share the same idea with Kim et al. (2020) and pays attention to the shift of attention to the knowledge between each turn of dialogue. The inspiration behind this is that the topic flows in conversation and so is the grounding of utterance. A knowledge shifter is devised to track the shift of attention on different knowledge. Apart from VAE-based methods, Zhao et al. (2020b) devise KnowledGPT, the state-of-the-art KGC model. It utilizes reinforcement learning for knowledge selection, whose reward is calculated as the similarity between the golden knowledge and the hypothesis by generator when fed with the context and the selected knowledge. After that, curriculum learning is used to jointly optimize the knowledge selector and generator in a unified framework.

Despite the abundance of common conversation data, human annotation is highly relied on to label the golden knowledge. Since human labor is expensive, a KGC model that requires little resource of annotated knowledge labels to train is of great significance. Zhao et al. (2020a) propose a disentangled model composed of a language model, a context processor and a document reader, which are conditionally independent and coordinated by a decoding manager. The language model is warmed up with conversation corpus without external knowledge and then all the components are trained on limited context-knowledge-response triples. Li et al. (2020) achieves zero-source KGC with no need for context-knowledge-response triplets. It introduces two latent variables to decide which piece of knowledge to select and how much the selected knowledge related to the response. The model is optimized with Generalized EM methods as well as a knowledge selection loss and mutual information loss. After trained on a cleaned version of reddit corpus, it accomplishes impressive results on two benchmarks in KGC.

### 4.2. Grounded by knowledge graph

However, the quality of unstructured knowledge is mixed, since they are usually directly collected from a website. Thus, some researchers propose to pre-process knowledge beforehand into the format of knowledge graph, composed of knowledge triplets (entity, relation, entity) or (entity, relation, item), which is much cleaner than unstructured knowledge sentences or documents.

Similar to the dialogue grounded by knowledge graph, this task requires a model to first retrieve a subgraph from a global commonsense knowledge graph, which is usually composed of the key entities and their one-hop neighbors (Zhou et al., 2018b; Zhang et al., 2019b). After that, to utilize one or more retrieved subgraphs, a model needs to integrate all the subgraphs into a highly abstract representation and help the generation process with the representation of all subgraphs.

Researchers have various ideas about how to construct and utilize knowledge in the format of a graph. Concretely, some establishes a common conversational graph from dialogue corpus as knowledge graph (Xu et al., 2020b), some pays attention to dialogue in a specific domain and establish a domain-specific knowledge graph. For example, Zhu et al. (2017) establish GenDS for music-related conversation and Liu

et al. (2018) construct a knowledge graph from Douban[2] to ground movie-related dialogues. Some grounds dialogue with a large-scale open-domain commonsense knowledge graph (Young et al., 2018). In this survey, we mainly focus on the last type and briefly introduce several representative methods.

CCM(Zhou et al., 2018b) is built on the basis of ConceptNet (Speer et al., 2017). It retrieves a subgrph for every word in context. If the word is a key concept, the subgraph is constituted of the concept node itself and its one-hop neighbors. If the word is a meaningless function word, a special subgraph will be returned. It devises a static graph attention and dynamic graph attention mechanism. The former encodes a subgraph with multiple triplets into a fixed-size vector, which is then concatenated with the word embedding for supplement. The latter is applied in the generation process in which the model interactively attends to knowledge triplets according to its decoder hidden state and decides to generate a generic word or copy an entity word. Concept Flow (Zhang et al., 2019b) extends the range of a subgraph to the two-hop neighbor of the concept node. The one-hop neighbors and the two-hop neighbors consist of the central graph and outer graph, respectively. The central graph is GraftNet (Sun et al., 2018), while the outer graph is processed in a similar way to static graph attention. Instead of fusing into the word embedding, the information in the knowledge graph is injected into the hidden state of the decoder. Wu et al. (2020a) make adaptation to the fusion of subgraphs. The authors claim that the subgraphs should not be treated equally and design a felicitous fact mechanism to select only useful subgraph. The response is also exploited to obtain the posterior probability of choosing every subgraph. In addition, to cope with the incompleteness in a graph, Tuan et al. (2019) focus on the zero-shot scenario and explores the possibility of update unseen knowledge graph with existing conversation data.

### 4.3. Grounded by persona

Dialogue grounded by user profile or persona is also referred to as personalized dialogue modeling. As the famous maxim by Shakespeare, there are a thousand Hamlet in a thousand people's eyes. Considering the profile of interlocutors help to promote the persona consistency of a dialogue system. As the name suggests, the core of personalized dialogue is how to utilize the user profiles or PERSONA. The task of personalized dialogue is to enable the dialogue system to mimic a specific persona. Given a dialogue context and a persona, the dialogue system should generate a response consistent with the designated persona.

Some early works like Li et al. (2016a) encode the profile of a speaker into a vector and fuses the profile vector into the decoder at every time step. Furthermore, the expression and tone of an interlocutor are also depended on the other participant. To cope with this, it proposes a speaker-addressee model, associating the persona of both speakers. Considering the scarcity of personalized data, Luan et al. (2017) take the advantage of multi-task learning and devises a new training method. Briefly, a seq2seq task is trained by general conversational data with a LSTM encoder-decoder architecture. An autoencoder task is trained to reconstruct the inputted persona with another LSTM encoder-decoder. The two decoders share parameters while performing multi-task learning, so the language model for generation could be adopted to the target speaker. Also to cope with data scarcity, Zhang et al. (2019c) devise a two-stage learning scheme named initialization-adaptation. The encoder is initialized with a large-scale general corpus and then fine-tuned with a small size of personalized training data. Innovatively, instead of using a special begin-of-sentence (BOS) token as the initial of the sentence, it trains a learning-to-start (LTS) model to predict the first word, which is believed by authors to boost the diversity.

A large body of works in personal dialogue leaves it to be self-evident that people are eager to show their persona intensively in dialogue. Zheng et al. (2020) doubt this assumption and believe that there is a persona-sparse issue: real-world dialogue data only contains a limited amount of utterance that correlates with the persona of a speaker. To tackle this, it revises the vanilla transformer architecture and devises an attention route mechanism to merge the persona information elastically. Song et al. (2020) is another work that adopts transformer to add new modules attending to persona text. Liu et al. (2020) take a new perspective and view personalized dialogue as an information perception process. It further proposes $\mathcal{P}^2$ BOT with a receiver that projects the impression and personas into a latent space and then measure their relevance as a reward for the training process.

Dissimilar to the works mentioned above, Li et al. (2021a) tackle the personalized dialogue from a different perspective. Instead of encoding the explicit profile into a vector, it gathers all the utterance history of a speaker as his/her implicit persona. The advantage of this method lies in its practicality since the utterance history is much easier to obtain compared with user profile. The personal utterance history serves as a personal mask to influence the context-response matching matrix.

### 4.4. Grounded by visual information

Human cognitive process is multi-modality in nature and bringing together vision and language in one intelligent conversational system has been one of the longest running goals of NLP (Shuster et al., 2020; Alamri et al., 2019). Grounding dialogue with an image or video clip is an intriguing and challenging task. In this section, we focus on grounded with image only and leave the survey for video clip grounding to be our future work. The task of vision-grounded dialogue is to talk on the background of an image. Given an image and a dialogue context, the dialogue system is supposed to give a response that is not only coherent in context, but also relate to the image.

Roughly, based on when to fuse the image information and textual context together, existing multi-modality methods fall into two categories: early fusion and late fusion (Shuster et al., 2020). Early-fusion methods extract a feature vector from an image and then the image feature vector is encoded with the text embedding together, usually by a transformer encoder or a RNN. On the contrary, late-fusion methods encode textual context independently before fusing the information in multi-modality.

Mostafazadeh et al. (2017) are the first to present a specific task of multi-modality dialogue, image-grounded conversation (IGC). Every conversation in the dataset is coupled with an image, serving as the background for the dialogue. Note that multi-modality dialogue is different from visual question answering (Antol et al., 2015) or visual dialogue (Das et al., 2017), for they do not concentrate on mimicking natural human conversation but focus on testing whether the model could comprehend an image or not. The proposed method is a typical late-fusion, processing textual information and vision information individually with a RNN encoder and a VGGnet. Feature vectors from two modalities are transformed into a joint initial hidden state for a decoder. Vision grounded dialogue has a wide range of applications in many vertical areas. Liao et al. (2018) introduce both image and external knowledge to ground dialogue in fashion and the e-economic area. To capture the multi-modal knowledge of fashion items mentioned in a conversation, it encodes style tips using representation obtained from the EI tree and injects the encoded knowledge into the initial hidden state of the decoder.

With the development of large-scale pre-trained vision model, directly taking the off-shelf representation from the ResNeXt or other pre-trained models seems a shortcut for multi-modality dialogue. Shuster et al. (2018) is one of them and is a late-fusion method. Its feature vector is extracted by ResNeXt (Xie et al., 2017). After a linear transformation, the feature vector of an image is appended at the end of context representation as a special token. Shuster et al. (2019) augment Shuster et al. (2018) with multi-task. Ideally, a dialogue system should be able to answer questions, ask questions, respond to statements,

[2] www.douban.com

provide useful information and knowledge. Motivated by this, it proposes dodecaDialogue, a set composed of 12 subtasks, measuring various properties for a dialogue system. Shuster et al. (2020) follow previous works and conducts experiments in both early fusion mode and late fusion mode. Specifically, in early fusion mode, the feature vector extracted from ResNeXt is concatenated with every word embedding as a visual supplement. Evaluation on COCO dataset (Lin et al., 2014) shows that early fusion mode results in a lower perplexity in some scenarios, perhaps due to the interaction between the image and dialogue.

To conclude, existing methods simply concatenate or append the feature vector of an image to the word embedding of dialogue text (early fusion) or to the representation of the dialogue text after encoder (late fusion). Sufficient and effective interaction between the dialogue and image is still a fruitful yet unexplored area.

## 5. Controllable: manage and interpret dialogue

Classic seq2seq generation model is notorious for lacking in interpretability. Toxic words and gender bias are also a hinder for the application of dialogue generation. To build a rational human-like conversation system, apart from external sources and groundings, it is of great significance to endow a dialogue system with the ability to control the dialogue strategy, expression style, or dialogue structure.

### 5.1. Control dialogue strategy

Social dialogue skills are necessary for a chatbot to give an interactive and engaging reply. Modeling human behavior and dialogue strategy enables the chatbot to capture the dialogue flow and therefore be more human-like. Therefore researching dialogue act and strategy is of great importance. In a nutshell, the task of controlling dialogue strategy is to maintain the dialogue appealing and meaningful by manipulating the trend and flow of dialogue.

Early paradigm of dialogue strategy is developed on dialogue game with a set of strategy heuristics (Yuan et al., 2007, 2008). In task-oriented dialogue, there is a specific goal like hotel booking or restaurant recommendation (Jeon and Lee, 2021; Henderson et al., 2019a). Thus the success or failure of a dialogue could be easily defined and observed. However, it is much difficult to define the result or state of open-domain dialogue, not to mention devising a mechanism to keep the dialogue going on. Despite its difficulty, a group of researchers have made efforts and various methods have been developed from vert different points of view (Li et al., 2016c; Yan and Zhao, 2018; Yu et al., 2016).

Li et al. (2016c) concentrate on stalemate detection and solution. It recognizes a stalemate in conversation with keyword matching of meaningless expression. Though sounds simple, the method works in a pragmatic way. Once the conversation falls into a stalemate, the system backtracks previous utterances and extract named entities to carry on the previous topic. Yan and Zhao (2018) regard the dialogue strategy of chatbot system as the ability to predict the unseen future utterance. It propounds a new paradigm that a dialogue system provides a (response, suggestion) pair given a query, believing that the suggestion can be used as a next utterance and bring information from an external scope. Wang et al. (2018a) study deepening or widening the chatting topics in a conversational session, or in other words, how to encourage the human user to talk more. It proposed a novel deep scheme featured with three channels to predict keywords, which could widen or deepen the topic of interest. Wang et al. (2018b) argue that putting up a new question is a useful strategy for a dialogue system, as a new question and topic could carry on the dialogue. It divides the whole vocabulary into three categories, namely interrogatives, ordinary words, and topic words, devising a hard-type-decoder and soft-type-decoder to dynamically determines the type of the next token.

Ke et al. (2018) propose to control dialogue strategy through three kinds of sentence functions, namely sentence function for interrogative response, imperative response, and declarative response. To generate a meaningful response with a sentence function, a latent variable subjecting to multivariate Gaussian distribution is introduced to project different sentence functions into different regions and decide dynamically the category of the next token in the decoding phase.

Other work models such social skills with several self-defined dialogue act policies or strategies. For example, Yu et al. (2016) propose 10 categorical states to describe the behavior of a human user like *Match Response*, *Don't Repeat*, *Ground on Named Entities*, etc. And reinforcement learning is used to perform strategy selection with a local and global reward. Different from previous works, Tang et al. (2019) raise a new task, target-guided open-domain dialogue. The chatbot is required to respond to the human speaker coherently, and proactively transmit the dialogue topic to finally reach a goal, which is a pre-defined keyword. At turn level, To avoid a curt shift in topic, a hybrid kernel-based method is responsible for predicting the keyword in the next utterance with K RBF kernels. At the discourse level, greedy search is applied to strictly approach the target topic word step by step.

### 5.2. Control dialogue style

Dialogue is diverse in nature as there exist various expression styles that could converse the same semantics, one of the major factors for the one-to-many relationship in dialogue. Hence steering the style of a dialogue model boosts the diversity of generation.

Formally, the task of stylized dialogue generation is to generate a response in the desired style, which could be different from the style of utterance history. Many methods have been tried to manipulate the style, such as adding a style embedding (Zheng et al., 2021), fusing the style information into the initial decoder state (Lample et al., 2018) or introducing latent variables to represent a dialogue style (Gao et al., 2019c).

Similar to knowledge-grounded dialogue, one major challenge for style transfer or stylistic generation is the paucity of parallel data in source style and target style. Thus, it is more practical to train a model with limited or no parallel data. Similar to Zhang et al. (2019c), Akama et al. (2017) first pre-train a seq2seq model on ordinary conversational corpus and then fine-tune the model with limited stylistic data. Besides, unsupervised learning and semi-supervised learning are proved to be effective methods in this subfield. Wu et al. (2020b) offer three paradigms for low resource stylized generation. That is, pivot-based, teacher-student and back-translation. The pivot-based method is a two-stage method that first generates a response in source style and then transfers it into target style. The teacher-student method train a style converter on parallel corpus as a teacher model, whose role is guiding a vanilla Seq2Seq model to generate in the desired style. The back-translation method trains a converter that transforms the text in target style into source style, therefore making use of unpaired corpus in a target style to construct pseudo label.

Yang et al. (2020) follow the teacher-student method in Wu et al. (2020b) and takes the advantage of large-scale pre-trained language model GPT-2 and DialoGPT. Specifically, a language model is trained with parallel stylized data on the basis of GPT-2. And a DialoGPT is trained with ordinary negative log-likelihood (NLL) loss on conversational data. The control over response style is realized with two training objectives. The word-level objective is to minimize the distance between the distribution given by the GPT-2 and the distribution given by the DialoGPT. The sentence-level objective is to maximize the confidence score given by a discriminative model, which is optimized via Gumbel trick.

Zheng et al. (2021) adopt the back-translation framework in Wu et al. (2020b) and regard the unpaired text in the desired style as a response for a lurking query and learns an inverse transformer network to predict the lurking query. Besides, style embedding is concatenated with the representation of context to control the style of generation. With limited parallel stylized data, existing methods tend to give

responses that are either stylized or either less context-relevant. Gao et al. (2019c) attribute this to the unbalanced distribution of latent variables. It observes that in Luan et al. (2017), data points in the latent space tend to form separate groups of clusters and fail to fill in the latent space.

Besides, controlling over expression style is regarded as an effective way to avoid rude and toxic expressions. Niu and Bansal (2018) focus on the politeness of response and comes up with a reinforcement learning model, exploiting the confidence of a politeness classifier as a part of the reward.

### 5.3. Control dialogue emotion

An intelligent chatbot should be able to perceive the emotion of a speaker and generate an emotional consistent response accordingly, mimicking behavior in human-human conversation like *empathy* and *comfort*. Emotional dialogue system could serve as an emotional companion and shows a promising prospect in psychological counseling.

Formally, the task of emotional dialogue requires a dialogue system to generate a response with a designated emotion label or a self-predicted emotion. The latter is also named *empathetic computing*, which is harder since a dialogue model needs to detect the appropriate emotion itself.

The exploration of affective text generation in a data-driven way starts with Affect-LM(Ghosh et al., 2017), the first work that explores how to generate emotional text, or affective generation. The backbone of this model is a vanilla LSTM structure, augmented by ann emotional vector updated at every step. The emotional vector is a discrete one-hot vector, with each dimension corresponding to a emotion category, calculated with linguistic inquiry and word count (LIWC). After Ghosh et al. (2017), Zhou and Wang (2017) take a radical step and collects Twitter conversations that include emojis, assuming that the emojis convey the underlying emotion of the speaker. The emotion of a speaker is modeled with a latent variable under the CVAE framework. Policy gradient method is also applied to give a high reward for an emotion-consistent response. Hu et al. (2017) also employ latent variables to capture the underlying emotional pattern in the text. Yet it claims that only a latent variable is not enough and is often agnostic to holistic features. Thus a structured code is introduced for independent attribute control, without entangling with other attributes, especially those implicitly modeled. The works mentioned above focus on affective text generation, building up the fundamental of emotional dialogue. And many techniques in emotional dialogue are borrowed from affective text generation.

Zhou et al. (2018a) are the first to notice the emotional factor in conversation generation and put forward the emotional dialogue problem, in which a dialogue system is required to generate an emotional response given a post and a pre-defined emotion category of response. To control the emotion in generated response, like Li et al. (2016a), each emotion is embedded into a real-valued, low dimensional vector. Apart from that, there are an internal emotional state and an external memory, which models the decaying of emotional state in decoding and chooses a token from an emotional-specific vocabulary. Asghar et al. (2018) is another work on affective response generation. Similar to Zhou et al. (2018a), it consists of three components: a vector embedding based on Valence/Arousal/Dominance dimensions (Mohammad, 2018), a training objective to maximize the affective consistency between the query and the response and finally an enhanced beam search decoding strategy. Following Asghar et al. (2018), Colombo et al. (2019) also project each word into a three-dimension vector using VAD lexicon (Mohammad, 2018). Besides, the traditional negative log-likelihood (NLL) loss function is augmented with an affection regularizer. The regularizer encourages the average VAD vector of generated hypothesis to get close to the VAD vector of the desired response. In decoding phrase, following Zhou et al. (2018a), the emotion of a whole sentence is imposed by every word position and decaying to zero step by step.

### 5.4. Control dialogue structure

Human conversation is highly structured. In a session of a multi-turn dialogue, the topic shift and the semantic correlation could be organized and described by a graph (Hu et al., 2019; Xu et al., 2020a) or other structural model (Zhai and Williams, 2014). Learning to construct dialogue structure yields basic insight into the structure of conversation and is believed to be an important step towards controllable and interpretable dialogue.

Early works in this topic deal with the problem through supervised training with human annotation (Jurafsky and Shriberg, 1997) or unsupervised training with Hidden Markov Model (HMM) (Zhai and Williams, 2014; Ritter et al., 2010). In recent years, with the proposal of VAE (Kingma and Welling, 2013), VRNN and VHRED are all applied to capture the structure of dialogue. However, these models mostly model the dialogue structure with a latent variable, which is implicit and little interpretable. Thus a group of researchers pay attention to human-readable dialogue structure, usually in the format of a graph. Though there exist studies of graph dialogue structure in task-oriented dialogue (Shi et al., 2019; Sun et al., 2021), little work sheds light on its counterpart in open-domain dialogue.

GSN(Hu et al., 2019) is the first to introduce graph-based representation learning into dialogue. The core of GSN is an utterance-level graph-based encoder. Every utterance is encoded into a vertex in a graph, whose representation is calculated with a RNN and a user matrix. The edges between the vertex are directed and determined by their topical order in a session. Its promising experiment result on Ubuntu Dialogue Corpus verifies the effectiveness of graph dialogue structure. DVAE-GNN (Xu et al., 2020a) extends the graph construction in GSN to a two-layer directed graph, which is helpful to capture holistic information in both utterance level and session level. The vertex in the lower layer represents an utterance, while the vertex in the upper layer represents a complete dialogue session. The vertex representation is calculated with a RNN encoder and a feed-forward net. The edges between the vertex are determined by their co-occurrence frequency in the corpora. A large portion of works limit their research in dyadic conversation. VRNN-LinearCRF(Qiu et al., 2020) therefore takes the multi-party conversation into consideration and comes up with an utterance-level dependency tree, with each path in the dependency tree corresponding to a thread in Ubuntu Dialogue Corpus (Lowe et al., 2015). The dependency tree is then utilized to detect the speaker-addressee relationship in multi-party dialogue. A linear CRF attention mechanism is also adopted to replace the traditional attention layer. Recently, Zhao et al, (2021) propose to control the structure style of knowldge-grounded dialogue with a novel variational segmentation method.

## 6. Evaluation metric

Evaluation for an open-domain conversation system could be performed manually or in an automatic way. Though human evaluation is reliable and consistent with user experience, human judgment is too expensive and time-consuming to apply on the full-volume test set. Thus automatic evaluation is also crucial for language generation tasks, especially in the early stage of model development. Here we center on automatic evaluation in this section.

For generation-based model, emphasis is put on appropriateness and informativeness. A variety of metrics are then devised or borrowed from other areas. BLEU (Papineni et al., 2002) computes the geometric mean of the precision for n-grams (n = 1,2,3,4). ROUGE(Lin, 2004) is a recall-oriented metric that focuses on the exact match of n-grams and ROUGE-L especially pays attention to the longest common sequence in hypothesis and generation. METEOR (Denkowski and Lavie, 2014) computes the harmonic mean of unigram precision and recall, featuring stemming and synonymy matching, which is rare in other metrics. BLEU, ROUGE and METEOR are all borrowed from machine translation.

Distinctness (Li et al., 2015) is calculated as the ratio of unique unigrams (Distinct-1) and bigrams (Distinct-2). It measures the diversity of generated text and is a good tool to detect safe and generic responses. Perplexity is calculated as the normalized inverse probability of the test set, or the exponential of the cross-entropy between the real language distribution and the language probability distribution learned by the model. Lower perplexity usually indicates the generated text is grammatically correct.

For retrieval-based model, most metrics are borrowed from information retrieval evaluation. Precision@n, mean average precision (MAP) and Normalized Discounted Cumulative Gain (nDCG) are popular metrics for a retrieval-based dialogue model.

Unfortunately, the automatic metrics sometimes disagree with human judgment, and some researchers find that there is little correlation between automatic evaluation and human judgment. And as pointed out in Gao et al. (2019b), these metrics are designed to measure the quality of a language model at corpus level, not sentence level. So it is worth discussing whether these metrics are really appropriate.

To tackle the shortcoming of these string-based automatic evaluation metrics, there is a surge in neural trainable metrics. As is indicated by its name, a neural trainable metric is trained with a neural network. For example, ADEM(Lowe et al., 2017) is trained with a variant of VHRED (Serban et al., 2016) model and gives a discrete score ranging from 1 to 5 to measure the quality of the generated text. RUBER(Tao et al., 2018) is an unsupervised learnable metric with an embedding-based scorer and a neural-based scorer. The former measures the similarity between the generated text and reference, while the latter focuses on the relatedness between the generated text and the query. Recently, with the rapid development of large-scale pre-trained language models, applying pre-trained language models to evaluate text generation is an intriguing topic. Sellam et al. (2020) propose to post-train BERT (Devlin et al., 2018) with five training objectives including back-translation likelihood and textual entailment. Experiment results on WMT Metrics Shared Task demonstrate its superiority over regular automatic metrics. Similarly, USR(Mehri and Eskenazi, 2020) is an unsupervised and reference-free method that utilizes Roberta (Liu et al., 2019) to calculate the log-likelihood of a response as a score. Besides, its chief novelty lies in using a retrieval model to evaluate the informativeness of a generated response based on the assumption that bland response usually is less probable to be retrieved.

However, neural trainable metrics are not perfect as well, whose defects have been discussed in Gao et al. (2019b).

## 7. Corpus and dataset

Recently, with the rapid development of internet, social medias and online forums are growing prosperously. The comment and communication online provides abundant resource to train a language model. There appears a great number of benchmark and dataset, recording the progress in every subfield in open-domain dialogue. In this section, we have a brief look at the some popular dataset and benchmark, mainly concentrating on their topic, source, language, data scale, and the design features. Basic statistics of these datasets are listed in Table 2.

**Short Text Conversation (STC)** (Shang et al., 2015) is a Chinese dialogue corpus extracted from Sina Weibo, a Twitter-like microblogging service in China. A Weibo user could post a short message whose length is no more than 140 words, and other users comment on the post, which is regarded as responses. There are 219,905 posts and 4,308,211 responses in the training data, precisely. On average, every post has 20 responses in the dataset, rendering it a good source for studying one-to-many in dialogue. Zhou et al. (2017) extend the dataset with a automatically annotated emotion label for emotional dialogue, though the accuracy of the emotion classifer is not perfect.

**Ubuntu Corpus** (Lowe et al., 2015) is an English corpus focusing on technical problems in Ubuntu system. Strictly speaking, it is like a task-specific corpus as the domain is confined in Ubuntu operating system. The corpus is collected from Ubuntu related chat room on Freenode Internet Relay Chat (IRC) and the interaction in chat room is filtered and simplified to dyadic conversation between two human users. It contains 930,000 dialogues and 7,100,000 utterances in total and is widely adopted in retrieval-based methods.

**Table 2**
Statistics of several frequently-used or newly-proposed dataset in various subfields.

| name | source | language | corpus statistics | feature |
|---|---|---|---|---|
| STC (Shang et al., 2015) | Weibo | Chinese | 219,905 posts, 4,308,211 responses | open-domain dialogue |
| Ubuntu Corpus (Lowe et al., 2015) | website | English | 930,000 dialogues, 7,100,000 utterances | open-domain dialogue |
| DailyDialog (Li et al., 2017b) | website | English | 13,118 dialogues | personalized dialogue |
| PERSONA-CHAT ( Zhang et al., 2018b) | crowd sourcing | English | 10,907 dialogues, 15,602 utterances | personalized dialogue |
| PersonalDialog (Zheng et al., 2019) | Weibo | Chinese | 8.48M persona, 20.83M dialogues | personalized dialogue |
| Wizard of Wikipedia ( Dinan et al., 2019) | crowd sourcing | English | 22,311 dialogues | KGC |
| Topical-Chat ( Gopalakrishnan et al., 2019) | crowd sourcing | English | 11,319 dialogues | KGC |
| EmpatheticDialogues ( Rashkin et al., 2018) | crowd sourcing | English | 32 emotion labels and 24,580 dialogues | empathetic computing |
| KdConv (Zhou et al., 2020) | crowd sourcing | Chinese | 4500 dialogues, 85,536 utterances | KGC |
| StickerChat (Gao et al., 2020) | social media | Chinese | 340,168 dialogues, 6,803,360 utterances | open-domain dialogue (with image stickers) |
| Image-Chat (Shuster et al., 2018) | crowd sourcing | English | 186,782 dialogues, 201,779 images | multi-modality dialogue (with style traits) |

**DailyDialog** (Li et al., 2017b) is a multi-turn dialogue dataset in English. Different from most dataset, the conversational data is crawled from a series of websites that serve for English beginners. So the conversation is human-written and thus of high quality. Specially, both the emotion and dialogue act are manually labeled for each utterance, with the emotion in 6 categories and dialogue act in 4 categories. However, the emotion label in the dataset is unbalanced in category, as there only exist about 5% utterances that have an emotion label other than “none” or “happy”. It contains 13,118 dialogues with 7.9 turns in a dialogue on average. It is a one-reference dataset, so Gupta et al. (2019) extend it with multi-reference.

**PERSONA-CHAT**(Zhang et al., 2018b) is a multi-turn dialogue dataset for emotional dialogue in English. It contains 1155 different personas from crowd sourcing, with 5 descriptive sentences for each persona. Speakers are designated with random persona and required to not only intensively show their own persona, but answer questions of their partner at the same time. There are 15,602 utterances over 10,907 dialogues provided in this dataset.

**PersonalDialog** (Zheng et al., 2019) is a Chinese dataset for personalized dialogue. Very similar to STC(Shang et al., 2015), it is collected from Weibo, but the scale is much larger thanks to the rapid

development of social media. Unlike PERSONA-CHAT, the persona in PersonalDialog is not designated ahead of time, but is implied in the public user profile, including gender, age, interest tag, and self-description. Besides, another appealing property of this dataset is equipped with multi-reference due to the intrinsic feature of Weibo. There are 8.47M persona, 20.83M dialogues and 56.25M utterances in total.

**Wizard of Wikipedia (WoW)** (Dinan et al., 2019) is an English dataset for knowledge-grounded conversation. All conversations are dyadic with two interlocutors, an apprentice and a wizard. The two participants are not quite symmetric: the wizard has access to an information retrieval system that shows some paragraphs from Wikipedia possibly relevant to the conversation. The dataset is widely adopted to verify the effectiveness of a KGC model. It contains 166,787 utterances and 1247 topics in training set, and the test set is divided into two parts: Test Seen and Test Unseen. The former only contains topics that appearing in the train set, while the Test Unseen has no such constraint.

**Topical-chat** (Gopalakrishnan et al., 2019) is another English dataset for knowledge-grounded conversation collected with Amazon Mechanical Turk. For every conversation, two human partners are provided with an article from Washington Post and knowledge about the three entities that most frequently appear in the article. The chief difference from the Wizard of Wikipedia (WoW) lies in the configuration of the dialogue setting. Apart from the ordinary wizard-apprentice setting, the researchers also explore other various patterns. They set 4 configurations in all and split the validation set and test set into 2 subsets respectively, with one containing entities frequently seen in training set and another containing entities that are rare seen or talked about, following Dinan et al. (2019).

**EmpatheticDialogues (ED)** (Rashkin et al., 2018) is proposed to facilitate evaluating models' ability to produce empathetic responses. Note that the goal of empathetic computing is different from aforementioned emotional dialogue. The former requires a model to be appropriate to signals purely inferred from text, while the latter asks the model to convey a pre-specific emotion in response. Each conversation is conducted by a listener and a speaker. The speaker first select an emotion label and then describe a situation when they feel that way, and then converse with the speaker based on the conceived situation. The dataset includes 32 emotion labels and 24,580 dialogues, with at least 4 utterances in a dialogue.

**KdConv** (Zhou et al., 2020) is Chinese dataset for knowledge-driven dialogue. All the dialogues are grounded by base knowledge graph in three domains: film, music and travel. Moreover, each utterance in a dialogue is annotated with a corresponding path in knowledge graph. Compared with other dataset, the average turn in a dialogue is remarkable with 19 turns in a dialogue on average. The dataset contains 4500 dialogues and 85,596 utterances in total.

**StickerChat** (Gao et al., 2020) is a large-scale multi-turn dialogue dataset with stickers. This dataset proposes the task of sticker response selection in multi-turn dialogue, where an appropriate sticker is recommended based on the dialogue history. In this dataset, all stickers are resized to a uniform size of $128 \times 128$ pixels. 20 utterances ahead of the sticker response are treated as the dialogue context, and irrelevant utterance sentences are filtered. After pre-processing, there are 320,168 context-sticker pairs in the training dataset, 10,000 pairs in the validation, and 10,000 pairs in test dataset respectively.

**Image-Chat** (Shuster et al., 2018) is a dataset for multi-modality dialogue, which is a large collection of (image, style trait for speaker A, style trait for speaker B, dialogue between A & B) in English. There are 215 style traits in all, categorized into three classes: positive, neural and negative. The dataset is collected from crowd-workers, who are required to talk about an image according to their designated style trait. The training set contains 186,782 dialogues and images in total, and every dialogues has 1.9 utterances on average.

## 8. Conclusion and open challenge

There is substantial literature on the task of open-domain dialogue, and recent years have witnessed rapid progress in this area. Yet open-domain dialogue is far from well-explored and it leaves several open challenges:

- **Logical Consistency** Existing methods mostly pay attention to semantics coherence, omitting the internal logic in dialogue. For example, a dialogue system may generate a contradictory response or exhibit an incompatible persona, especially in a multi-turn dialogue. Maybe the incorporation of logic structure in dialogue game and computational dialectics (Yuan et al., 2011; Mackenzie, 1990) is a hopeful remedy. Till now, self-consistence of a dialogue model is still an open problem.
- **Interpretability and Controllability** There is a lack in interpretability for existing methods, as the generation is usually a neural probabilistic process. Poor interpretability leads to poor controllability to some extent, though some work avoids toxic words or unethical responses with reinforcement learning (Niu and Bansal, 2018) or curriculum learning (Shen and Feng, 2020).
- **Efficiency and Compactness** The appearance of large-scale pre-trained language models facilitates training of downstream task at a great extent. However, a model with a complex structure and a large size of parameters is often time-consuming when training and inference. And a specific downstream task often requires a new set of parameters. multi-task learning and continual learning seem to be solutions for this, and more efforts need to be made towards the efficiency and compactness of a dialogue model.

In this paper, we formulate three frameworks in open-domain dialogue and conclude two long running goals for neural open-domain dialogue as well as the efforts to advance towards these goals. We review several metrics for evaluating the quality of a dialogue system and list several frequently-used corpus and benchmarks for our readers.

## Declaration of competing interest

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

## References

Adiwardana, D., Luong, M.T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al., 2020. Towards a Human-like Open-Domain Chatbot arXiv preprint arXiv:2001.09977.

Akama, R., Inada, K., Inoue, N., Kobayashi, S., Inui, K., 2017. Generating stylistically consistent dialog responses with transfer learning. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, ume 2, pp. 408–412. Short Papers).

Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., et al., 2019. Audio visual scene-aware dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7558–7567.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D., 2015. Vqa: visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433.

Asghar, N., Poupart, P., Hoey, J., Jiang, X., Mou, L., 2018. Affective neural response generation. In: European Conference on Information Retrieval. Springer, pp. 154–166.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR.

Bao, S., He, H., Wang, F., Wu, H., 2019. Plato: Pre-trained Dialogue Generation Model with Discrete Latent Variable arXiv preprint arXiv:1910.07931.

Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S., 2015. Generating Sentences from a Continuous Space arXiv preprint arXiv:1511.06349.

Cai, D., Wang, Y., Bi, V., Tu, Z., Liu, X., Lam, W., Shi, S., 2018. Skeleton-to-response: Dialogue Generation Guided by Retrieval Memory arXiv preprint arXiv:1809.05296.

Chen, Q., Wang, W., 2019. Sequential Attention-Based Network for Noetic End-To-End Response Selection arXiv preprint arXiv:1901.02609.

Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2020. Electra: Pre-training Text Encoders as Discriminators rather than Generators arXiv preprint arXiv:2003.10555.
Colombo, P., Witon, W., Modi, A., Kennedy, J., Kapadia, M., 2019. Affect-driven Dialog Generation arXiv preprint arXiv:1904.02793.
Daniel, J., James, H.M., 2000. Speech and Language Processing.
Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D., 2017. Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 326–335.
Denkowski, M., Lavie, A., 2014. Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation.
Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv preprint arXiv: 1810.04805.
Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J., 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. ICLR.
Feng, S., Chen, H., Li, K., Yin, D., 2020. Posterior-gan: towards informative and coherent response generation with posterior generative adversarial network. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7708–7715.
Gao, J., Bi, W., Liu, X., Li, J., Shi, S., 2019a. Generating multiple diverse responses for short-text conversation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6383–6390.
Gao, J., Galley, M., Li, L., 2019b. Neural approaches to conversational AI: question answering, task-oriented dialogues and social chatbots. In: Now Foundations and Trends.
Gao, S., Chen, X., Liu, C., Liu, L., Zhao, D., Yan, R., Gao, S., Chen, X., Liu, C., Liu, L., Zhao, D., Yan, R., 2020. Learning to Respond with Stickers: A Framework of Unifying Multi-Modality in Multi-Turn Dialog. WWW.
Gao, X., Zhang, Y., Lee, S., Galley, M., Brockett, C., Gao, J., Dolan, B., 2019c. Structuring Latent Spaces for Stylized Response Generation arXiv preprint arXiv:1909.05361.
Ghazvininejad, M., Brockett, C., Chang, M.W., Dolan, B., Gao, J., Yih, W.t., Galley, M., 2018. A Knowledge-Grounded Neural Conversation Model. AAAI.
Ghosh, S., Chollet, M., Laksana, E., Morency, L.P., Scherer, S., 2017. Affect-lm: A Neural Language Model for Customizable Affective Text Generation arXiv preprint arXiv: 1704.06851.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.
Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., Hakkani-Tür, D., Ai, A.A., 2019. Topical-chat: towards knowledge-grounded open-domain conversations. Proc. Interspeech 2019, 1891–1895.
Gu, X., Cho, K., Ha, J.W., Kim, S., 2018. Dialogwae: Multimodal Response Generation with Conditional Wasserstein Auto-Encoder arXiv preprint arXiv:1805.12352.
Gupta, P., Mehri, S., Zhao, T., Pavel, A., Eskenazi, M., Bigham, J.P., 2019. Investigating Evaluation of Open-Domain Dialogue Systems with Human Generated Multiple References arXiv preprint arXiv:1907.10568.
Han, J., Hong, T., Kim, B., Ko, Y., Seo, J., 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1549–1558.
Henderson, M., Vulić, I., Casanueva, I., Budzianowski, P., Gerz, D., Coope, S., Spithourakis, G., Wen, T.H., Mrkšić, N., Su, P.H., 2019a. Polyresponse: A Rank-Based Approach to Task-Oriented Dialogue with Application in Restaurant Search and Booking arXiv preprint arXiv:1909.01296.
Henderson, M., Vulić, I., Gerz, D., Casanueva, I., Budzianowski, P., Coope, S., Spithourakis, G., Wen, T.H., Mrkšić, N., Su, P.H., 2019b. Training Neural Response Selection for Task-Oriented Dialogue Systems arXiv preprint arXiv:1906.01543.
Hsueh, P.Y., Moore, J.D., Renals, S., 2006. Automatic segmentation of multiparty dialogue. In: 11th Conference of the European Chapter of the Association for Computational Linguistics.
Hu, B., Lu, Z., Li, H., Chen, Q., 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. NIPS, pp. 2042–2050.
Hu, W., Chan, Z., Liu, B., Zhao, D., Ma, J., Yan, R., 2019. Gsn: A Graph-Structured Network for Multi-Party Dialogues arXiv preprint arXiv:1905.13637.
Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., Xing, E.P., 2017. Toward controlled generation of text. In: International Conference on Machine Learning. PMLR, pp. 1587–1596.
Huang, M., Zhu, X., Gao, J., 2020. Challenges in building intelligent open-domain dialog systems. ACM Trans. Inf. Syst. 38, 1–32.
Jeon, H., Lee, G.G., 2021. Dora: toward Policy Optimization for Task-Oriented Dialogue System with Efficient Context arXiv preprint arXiv:2107.03286.
Jurafsky, D., Shriberg, E., 1997. Switchboard Swbd-Damsl Shallow-Discourse-Function Annotation Coders Manual, Draft 13 Daniel Jurafsky*, Elizabeth Shriberg+, and Debra Biasca** University of colorado at Boulder &+ Sri International.
Kadlec, R., Schmid, M., Kleindienst, J., 2015. Improved Deep Learning Baselines for Ubuntu Corpus Dialogs arXiv preprint arXiv:1510.03753.
Kaiser, L., Bengio, S., Roy, A., Vaswani, A., Parmar, N., Uszkoreit, J., Shazeer, N., 2018. Fast decoding in sequence models using discrete latent variables. In: International Conference on Machine Learning. PMLR, pp. 2390–2399.
Kang, L., Hu, B., Wu, X., Chen, Q., He, Y., 2014. A short texts matching method using shallow features and deep features. In: CCF International Conference on Natural Language Processing and Chinese Computing. Springer, pp. 150–159.
Ke, P., Guan, J., Huang, M., Zhu, X., 2018. Generating informative responses with controlled sentence function. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ume 1, pp. 1499–1508. Long Papers).
Kim, B., Ahn, J., Kim, G., 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue, 07510 arXiv preprint arXiv:2002.
Kingma, D.P., Welling, M., 2013. Auto-encoding Variational Bayes arXiv preprint arXiv: 1312.6114.
Lample, G., Subramanian, S., Smith, E., Denoyer, L., Ranzato, M., Boureau, Y.L., 2018. Multiple-attribute text rewriting. In: International Conference on Learning Representations.
Lee, J., Mansimov, E., Cho, K., 2018. Deterministic Non-autoregressive Neural Sequence Modeling by Iterative Refinement arXiv preprint arXiv:1802.06901.
Lewin, I., Lane, M., 2000. A formal model of conversational game theory. In: Proceedings of the 4th Workshop on the Semantics and Pragmatics of Dialogue (Gotalog). Citeseer.
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020a. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880.
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al., 2020b. Retrieval-augmented Generation for Knowledge-Intensive Nlp Tasks arXiv preprint arXiv:2005.11401.
Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., 2015. A diversity-promoting objective function for neural conversation models. NAACL 110–119.
Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., Dolan, B., 2016a. A Persona-Based Neural Conversation Model. ACL, pp. 994–1003.
Li, J., Liu, C., Tao, C., Chan, Z., Zhao, D., Zhang, M., Yan, R., 2021a. Dialogue History Matters! Personalized Response Selectionin Multi-Turn Retrieval-Based Chatbots arXiv preprint arXiv:2103.09534.
Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., Gao, J., 2016b. Deep Reinforcement Learning for Dialogue Generation. EMNLP, pp. 1192–1202.
Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., Jurafsky, D., 2017a. Adversarial Learning for Neural Dialogue Generation. EMNLP, pp. 2157–2169.
Li, L., Xu, C., Wu, W., Zhao, Y., Zhao, X., Tao, C., 2020. Zero-resource Knowledge-Grounded Dialogue Generation arXiv preprint arXiv:2008.12918.
Li, X., Mou, L., Yan, R., Zhang, M., 2016c. Stalematebreaker: A Proactive Content-Introducing Approach to Automatic Human-Computer Conversation arXiv preprint arXiv:1604.04358.
Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S., 2017b. Dailydialog: A Manually Labelled Multi-Turn Dialogue Dataset arXiv preprint arXiv:1710.03957.
Li, Y., Xu, C., Hu, H., Sha, L., Zhang, Y., Jiang, D., 2021b. Small Changes Make Big Differences: Improving Multi-Turn Response Selection\\\\in Dialogue Systems via Fine-Grained Contrastive Learning arXiv preprint arXiv:2111.10154.
Li, Z., Niu, C., Meng, F., Feng, Y., Li, Q., Zhou, J., 2019. Incremental transformer with deliberation decoder for document grounded conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 12–21.
Lian, R., Xie, M., Wang, F., Peng, J., Wu, H., 2019. Learning to Select Knowledge for Response Generation in Dialog Systems arXiv preprint arXiv:1902.04911.
Liao, L., Ma, Y., He, X., Hong, R., Chua, T.s., 2018. Knowledge-aware multimodal dialogue systems. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 801–809.
Lin, C.Y., 2004. Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81.
Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.
Liu, Q., Chen, Y., Chen, B., Lou, J.G., Chen, Z., Zhou, B., Zhang, D., 2020. You Impress Me: Dialogue Generation via Mutual Persona Perception arXiv preprint arXiv: 2004.05388.
Liu, S., Chen, H., Ren, Z., Feng, Y., Liu, Q., Yin, D., 2018. Knowledge diffusion for neural dialogue generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol 1, pp. 1489–1498. Long Papers).
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A Robustly Optimized Bert Pretraining Approach arXiv preprint arXiv:1907.11692.
Lowe, R., Noseworthy, M., Serban, I.V., 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, 1, pp. 1116–1126. Long Papers).
Lowe, R., Pow, N., Serban, I., Pineau, J., 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. SIGDIAL, pp. 285–294.
Lu, Z., Li, H., 2013. A Deep Architecture for Matching Short Texts. NIPS, pp. 1367–1375.
Luan, Y., Brockett, C., Dolan, B., Gao, J., Galley, M., 2017. Multi-task Learning for Speaker-Role Adaptation in Neural Conversation Models arXiv preprint arXiv: 1710.07388.
Mackenzie, J., 1990. Four dialogue systems. Stud. Logica 49, 567–583.
McBurney, P., Parsons, S., 2003. Dialogue game protocols. In: Communication in Multiagent Systems. Springer, pp. 269–283.
Mehri, S., Eskenazi, M., 2020. Usr: an Unsupervised and Reference Free Evaluation Metric for Dialog Generation arXiv preprint arXiv:2005.00456.
Meng, C., Ren, P., Chen, Z., Ren, Z., Xi, T., Rijke, M.d., 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 522–532.
Meng, C., Ren, P., Chen, Z., Sun, W., Ren, Z., Tu, Z., Rijke, M.d., 2020. Dukenet: a dual knowledge interaction network for knowledge-grounded conversation. In:

Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1151–1160.
Miller, A., Fisch, A., Dodge, J., Karimi, A.H., Bordes, A., Weston, J., 2016. Key-value Memory Networks for Directly Reading Documents arXiv preprint arXiv: 1606.03126.
Mohammad, S., 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ume 1, pp. 174–184. Long Papers).
Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G.P., Vanderwende, L., 2017. Image-grounded Conversations: Multimodal Context for Natural Question and Response Generation arXiv preprint arXiv:1701.08251.
Niu, T., Bansal, M., 2018. Polite dialogue generation without parallel data. In: Transactions of the Association for Computational Linguistics, 6, pp. 373–389.
Pandey, G., Contractor, D., Kumar, V., Joshi, S., 2018. Exemplar encoder-decoder for neural conversation generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ume 1, pp. 1329–1338. Long Papers).
Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 311–318.
Qin, J., Ye, Z., Tang, J., Liang, X., 2020. April). Dynamic knowledge routing network for target-guided open-domain conversation (Vol. 34, No. 05,, 8657–8664.
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. URL: https://www.aclweb.org/anthology/P19-1372.
Qiu, L., Zhao, Y., Shi, W., Liang, Y., Shi, F., Yuan, T., Yu, Z., Zhu, S.C., 2020. Structured Attention for Unsupervised Dialogue Structure Induction arXiv preprint arXiv: 2009.08552.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language Models Are Unsupervised Multitask Learners.
Rakhlin, A., 2016. Convolutional Neural Networks for Sentence Classification. GitHub.
Rashkin, H., Smith, E.M., Li, M., Boureau, Y.L., 2018. Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset arXiv preprint arXiv: 1811.00207.
Ritter, A., Cherry, C., Dolan, W.B., 2010. Unsupervised modeling of twitter conversations. In: Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 172–180.
Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E.M., et al., 2020. Recipes for Building an Open-Domain Chatbot arXiv preprint arXiv:2004.13637.
Saleh, A., Baly, R., Barrón-Cedeno, A., Martino, G.D.S., Mohtarami, M., Nakov, P., Glass, J., 2019. Team Qcri-Mit at Semeval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection arXiv preprint arXiv:1904.03513.
Saleh, A., Jaques, N., Ghandeharioun, A., Shen, J., Picard, R., 2020. Hierarchical reinforcement learning for open-domain dialog. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8741–8748.
See, A., Roller, S., Kiela, D., Weston, J., 2019. What Makes a Good Conversation? How Controllable Attributes Affect Human Judgments arXiv preprint arXiv:1902.08654.
Sellam, T., Das, D., Parikh, A.P., 2020. Bleurt: Learning Robust Metrics for Text Generation arXiv preprint arXiv:2004.04696.
Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J., 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. AAAI, pp. 3776–3784.
Serban, I.V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A.C., Bengio, Y., 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. AAAI, pp. 3295–3301.
Shang, L., Lu, Z., Li, H., 2015. Neural Responding Machine for Short-Text Conversation. ACL, pp. 1577–1586.
Shao, H., Wang, J., Lin, H., Zhang, X., Zhang, A., Ji, H., Abdelzaher, T., 2021. Controllable and diverse text generation in e-commerce. In: Proceedings of the Web Conference 2021, pp. 2392–2401.
Shen, L., Feng, Y., 2020. Cdl: Curriculum Dual Learning for Emotion-Controllable Response Generation arXiv preprint arXiv:2005.00329.
Shen, X., Su, H., Li, Y., Li, W., Niu, S., Zhao, Y., Aizawa, A., Long, G., 2017. A Conditional Variational Framework for Dialog Generation arXiv preprint arXiv:1705.00316.
Shi, W., Zhao, T., Yu, Z., 2019. Unsupervised Dialog Structure Learning arXiv preprint arXiv:1904.03736.
Shin, J., Kim, Y., Yoon, S., Jung, K., 2018. Contextual-cnn: a novel architecture capturing unified meaning for sentence classification. In: 2018 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, pp. 491–494.
Shum, H., He, X., Li, D., 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. Front. IT & EE 19, 10–26.
Shuster, K., Humeau, S., Bordes, A., Weston, J., 2018. Image Chat: Engaging Grounded Conversations arXiv preprint arXiv:1811.00945.
Shuster, K., Ju, D., Roller, S., Dinan, E., Boureau, Y.L., Weston, J., 2019. The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents arXiv preprint arXiv:1911.03768.
Shuster, K., Smith, E.M., Ju, D., Weston, J., 2020. Multi-modal Open-Domain Dialogue arXiv preprint arXiv:2010.01082.
Song, H., Wang, Y., Zhang, W.N., Liu, X., Liu, T., 2020. Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation, 07672 arXiv preprint arXiv:2004.
Song, Y., Yan, R., Li, C.T., Nie, J.Y., Zhang, M., Zhao, D., 2018. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. IJCAI, pp. 4382–4388.
Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., Nie, J.Y., 2015a. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. CIKM, pp. 553–562.
Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, B., 2015b. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. NAACL, pp. 196–205.
Speer, R., Chin, J., Havasi, C., 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In: Thirty-first AAAI Conference on Artificial Intelligence.
Sun, B., Li, K., 2021. Neural dialogue generation methods in open domain: a survey. Nat. Lang. Process. Res. 1, 56–70.
Sun, H., Dhingra, B., Zaheer, M., Mazaitis, K., Salakhutdinov, R., Cohen, W., 2018. Open domain question answering using early fusion of knowledge bases and text. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp. 4231–4242. https://doi.org/10.18653/v1/D18-1455. URL: https://aclanthology.org/D18-1455.
Sun, Y., Shan, Y., Tang, C., Hu, Y., Dai, Y., Yu, J., Sun, J., Huang, F., Si, L., 2021. Unsupervised Learning of Deterministic Dialogue Structure with Edge Graph Auto-Encoder.
Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y., 2000. Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems, pp. 1057–1063.
Tang, J., Zhao, T., Xiong, C., Liang, X., Xing, E.P., Hu, Z., 2019. Target-guided Open-Domain Conversation arXiv preprint arXiv:1905.11553.
Tao, C., Feng, J., Yan, R., Wu, W., Jiang, D., 2021. January). A survey on response selection for retrieval-based dialogues (Vol. 21,, 4619–4626.
Tao, C., Mou, L., Zhao, D., Yan, R., 2018. Ruber: an unsupervised method for automatic evaluation of open-domain dialog systems. In: Thirty-Second AAAI Conference on Artificial Intelligence.
Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D., Yan, R., 2019. Multi-representation Fusion Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. WSDM, pp. 267–275.
Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D., & Yan, R. (2019, July). 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* (pp. 1-11).
Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., Zhao, D., 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol 2, pp. 231–236. Short Papers).
Tuan, Y.L., Chen, Y.N., Lee, H.y., 2019. Dykgchat: benchmarking dialogue generation grounding on dynamic knowledge graphs. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP), pp. 1855–1865.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. In: Attention Is All You Need. NIPS, pp. 5998–6008.
Wang, M., Lu, Z., Li, H., Liu, Q., 2015. Syntax-based Deep Matching of Short Texts. IJCAI, pp. 1354–1361.
Wang, W., Huang, M., Xu, X.S., Shen, F., Nie, L., 2018a. Chat more: deepening and widening the chatting topic via a deep model, in: Collins-Thompson, K., Mei, Q., 0001, B.D.D., 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. 255–264. URL: http://doi.acm.org/10.1145/3209978.3210061, doi:10.1145/3209978.3210061.
Wang, Y., Liu, C., Huang, M., Nie, L., 2018b. Learning to ask questions in open-domain conversational systems with typed decoders. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol 1, pp. 2193–2203. Long Papers).
Wang, Z., Li, S., Chen, G., Lin, Z., 2017. Deep and shallow features learning for short texts matching. In: 2017 International Conference on Progress in Informatics and Computing (PIC). IEEE, pp. 51–55.
Weston, J., Chopra, S., Bordes, A., 2014. Memory Networks arXiv preprint arXiv: 1410.3916.
Weston, J., Dinan, E., Miller, A.H., 2018. Retrieve and Refine: Improved Sequence Generation Models for Dialogue arXiv preprint arXiv:1808.04776.
Whang, T., Lee, D., Lee, C., Yang, K., Oh, D., Lim, H., 2020. An Effective Domain Adaptive Post-training Method for Bert in Response Selection. INTERSPEECH, pp. 1585–1589.
Whang, T., Lee, D., Oh, D., Lee, C., Han, K., Lee, D.h., Lee, S., 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14041–14049.
Wu, S., Li, Y., Zhang, D., Zhou, Y., Wu, Z., 2020a. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5811–5820.
Wu, Y., Wang, Y., Liu, S., 2020b. A dataset for low-resource stylized sequence-to-sequence generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9290–9297.
Wu, Y., Wei, F., Huang, S., Wang, Y., Li, Z., Zhou, M., 2019. Response generation by context-aware prototype editing. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7281–7288.

Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z., 2016. Sequential Matching Network: A New Architecture for Multi-Turn Response Selection in Retrieval-Based Chatbots arXiv preprint arXiv:1612.01627.
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. ACL, pp. 496–505.
Xia, Y., Tian, F., Wu, L., Lin, J., Qin, T., Yu, N., Liu, T.Y., 2017. Deliberation networks: sequence generation beyond one-pass decoding. Adv. Neural Inf. Process. Syst. 30, 1784–1794.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500.
Xing, C., Wu, W., Wu, Y., Zhou, M., Huang, Y., Ma, W.Y., 2017. Hierarchical Recurrent Attention Network for Response Generation arXiv preprint arXiv:1701.07149.
Xu, J., Lei, Z., Wang, H., Niu, Z.Y., Wu, H., Che, W., Liu, T., 2020a. Discovering Dialog Structure Graph for Open-Domain Dialog Generation arXiv preprint arXiv: 2012.15543.
Xu, J., Ren, X., Lin, J., Sun, X., 2018. Diversity-promoting gan: a cross-entropy based generative adversarial network for diversified text generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3940–3949.
Xu, J., Wang, H., Niu, Z.Y., Wu, H., Che, W., Liu, T., 2020b. Conversational graph grounded policy learning for open-domain conversation generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1835–1845.
Xu, R., Tao, C., Jiang, D., Zhao, X., Zhao, D., Yan, R., 2020c. Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-Based Dialogues arXiv preprint arXiv:2009.06265.
Xu, Y., Zhao, H., Zhang, Z., 2021. Topic-aware multi-turn dialogue modeling. In: The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21).
Xu, Z., Liu, B., Wang, B., Sun, C.J., Wang, X., Wang, Z., Qi, C., 2017. Neural response generation via gan with an approximate embedding layer. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 617–626.
Yan, R., 2018. Chitty-Chitty-Chat Bot": Deep Learning for Conversational Ai. IJCAI, pp. 5520–5526.
Yan, R., Song, Y., Wu, H., 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. SIGIR, pp. 55–64.
Yan, R., Zhao, D., 2018. Smarter Response with Proactive Suggestion: A New Generative Neural Conversation Paradigm. IJCAI, pp. 4525–4531.
Yang, L., Qiu, M., Qu, C., Guo, J., Zhang, Y., Croft, W.B., Huang, J., Chen, H., 2018a. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In: The 41st International Acm Sigir Conference on Research & Development in Information Retrieval, pp. 245–254.
Yang, M., Tu, W., Qu, Q., Zhao, Z., Chen, X., Zhu, J., 2018b. Personalized response generation by dual-learning based domain adaptation. Neural Network. 103, 72–82.
Yang, Z., Wu, W., Xu, C., Liang, X., Bai, J., Wang, L., Wang, W., Li, Z., 2020. Styledgpt: Stylized Response Generation with Pre-trained Language Models arXiv preprint arXiv:2010.02569.
Yao, K., Zweig, G., Peng, B., 2015. Attention with Intention for a Neural Network Conversation Model arXiv preprint arXiv:1510.08565.
Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., Huang, M., 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In: Thirty-Second AAAI Conference on Artificial Intelligence.
Yu, Z., Xu, Z., Black, A.W., Rudnicky, A., 2016. Strategy and policy learning for non-task-oriented conversational systems. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 404–412.
Yuan, C., Zhou, W., Li, M., Lv, S., Zhu, F., Han, J., Hu, S., 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. 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), pp. 111–120.
Yuan, T., Moore, D., Grierson, A., 2007. A human–computer debating system prototype and its dialogue strategies. Int. J. Intell. Syst. 22, 133–156.
Yuan, T., Moore, D., Grierson, A., 2008. A human-computer dialogue system for educational debate: a computational dialectics approach. Int. J. Artif. Intell. Educ. 18, 3–26.
Yuan, T., Moore, D., Reed, C., Ravenscroft, A., Maudet, N., 2011. Informal logic dialogue games in human–computer dialogue. Knowl. Eng. Rev. 26, 159–174.
Zhai, K., Williams, J.D., 2014. Discovering latent structure in task-oriented dialogues. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol 1, pp. 36–46. Long Papers).
Zhang, H., Lan, Y., Guo, J., Xu, J., Cheng, X., 2018a. Reinforcing Coherence for Sequence to Sequence Model in Dialogue Generation. IJCAI, pp. 4567–4573.
Zhang, H., Lan, Y., Pang, L., Guo, J., Cheng, X., 2019a. 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, pp. 3721–3730.
Zhang, H., Liu, Z., Xiong, C., Liu, Z., 2019b. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs arXiv preprint arXiv: 1911.02707.
Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J., 2018b. Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too?, 07243 arXiv preprint arXiv: 1801.
Zhang, W.N., Zhu, Q., Wang, Y., Zhao, Y., Liu, T., 2019c. Neural personalized response generation as domain adaptation. World Wide Web 22, 1427–1446.
Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., Dolan, B., 2018c. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization arXiv preprint arXiv:1809.05972.
Zhang, Y., Sun, S., Galley, M., Chen, Y.C., Brockett, C., Gao, X., Gao, J., Liu, J., Dolan, B., 2019d. Dialogpt: Large-Scale Generative Pre-training for Conversational Response Generation arXiv preprint arXiv:1911.00536.
Zhang, Z., Li, J., Zhu, P., Zhao, H., Liu, G., 2018d. Modeling Multi-Turn Conversation with Deep Utterance Aggregation. ACL, pp. 3740–3752.
Zhang, Z., Zheng, D., Gong, P., 2021. Multi-turn response selection in retrieval based chatbots with hierarchical residual matching network. In: Journal of Physics: Conference Series. IOP Publishing., 012023.
Zhao, T., Lee, K., Eskenazi, M., 2018. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation arXiv preprint arXiv: 1804.08069.
Zhao, T., Zhao, R., Eskenazi, M., 2017. Learning Discourse-Level Diversity for Neural Dialog Models Using Conditional Variational Autoencoders. ACL, pp. 654–664.
Zhao, X., Wu, W., Tao, C., Xu, C., Zhao, D., Yan, R., 2020a. Low-resource Knowledge-Grounded Dialogue Generation arXiv preprint arXiv:2002.10348.
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), pp. 3377–3390.
Zheng, Y., Chen, G., Huang, M., Liu, S., Zhu, X., 2019. Personalized Dialogue Generation with Diversified Traits arXiv preprint arXiv:1901.09672.
Zheng, Y., Chen, Z., Zhang, R., Huang, S., Mao, X., Huang, M., 2021. Stylized dialogue response generation using stylized unpaired texts. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14558–14567.
Zheng, Y., Zhang, R., Huang, M., Mao, X., 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9693–9700.
Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B., 2017. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory arXiv preprint arXiv:1704.01074.
Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B., 2018a. Emotional chatting machine: emotional conversation generation with internal and external memory. In: Proceedings of the AAAI Conference on Artificial Intelligence.
Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X., 2018b. Commonsense Knowledge Aware Conversation Generation with Graph Attention. IJCAI, pp. 4623–4629.
Zhou, H., Zheng, C., Huang, K., Huang, M., Zhu, X., 2020. Kdconv: A Chinese Multi-Domain Dialogue Dataset towards Multi-Turn Knowledge-Driven Conversation arXiv preprint arXiv:2004.04100.
Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., Liu, X., Yan, R., 2016. Multi-view Response Selection for Human-Computer Conversation. EMNLP, pp. 372–381.
Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W.X., Yu, D., Wu, H., 2018c. Multi-turn Response Selection for Chatbots with Deep Attention Matching Network. ACL, pp. 1118–1127.
Zhou, X., Wang, W.Y., 2017. Mojitalk: Generating Emotional Responses at Scale arXiv preprint arXiv:1711.04090.
Zhu, W., Mo, K., Zhang, Y., Zhu, Z., Peng, X., Yang, Q., 2017. Flexible End-To-End Dialogue System for Knowledge Grounded Conversation arXiv preprint arXiv: 1709.04264.
Zhao X, Tao C, Wu W, et al. A document-grounded matching network for response selection in retrieval-based chatbots[J]. arXiv preprint arXiv:1906.04362, 2019.
Zhao X, Fu T, Tao C, et al. Learning to Express in Knowledge-Grounded Conversation[J]. 2021.