ATLANTIS PRESS

Natural Language Processing Research
Vol. 1(3-4), 2021, *pp.* 56–70
DOI: https://doi.org/10.2991/nlpr.d.210223.001; eISSN: 2666-0512
https://www.atlantis-press.com/journals/nlpr/

Research Article

Neural Dialogue Generation Methods in Open Domain: A Survey

Bin Sun, Kan Li[*]

School of Computer Science, Beijing Institute of Technology, Beijing, 100081, China

ARTICLE INFO

Article History

Received 07 Dec 2020
Accepted 17 Feb 2021

Keywords

Neural dialogue generation
Open-domain dialogue system
Natural language processing

ABSTRACT

Open-Domain Dialogue Generation (human–computer interaction) is an important issue in the field of Natural Language Processing (NLP). Because of the improvement of deep learning techniques, a large number of neural dialogue generative methods were proposed to generate better responses. In this survey, we elaborated the research history of these existing generative methods, and then roughly divided them into six categories, *i.e.*, Encoder-Decoder framework-based methods, Hierarchical Recurrent Encoder-Decoder (HRED)-based methods, Variational Autoencoder (VAE)-based methods, Reinforcement Learning (RL)- based methods, Generative Adversarial Network (GAN)-based methods, and pretraining-model-based methods. We dived into the methods of each category and gave the detailed discussions of these methods. After that, we presented a comparison among the different categories of methods and analyzed their advantages and disadvantages. We enumerated some open access public datasets and some commonly used automatic evaluating metrics. Finally, we discuss some possible research directions that can take the research of neural dialogue generation into a new frontier in the future.

1. INTRODUCTION

The study of the dialogue system can be traced to the Turing test in 1950 [1]. If a machine can talk to humans without being able to identify its machine identity, then this machine is said to be intelligent. In other words, the development of automatic dialogue systems can reflect the development degree of artificial intelligence to a certain degree. Therefore, the dialogue system has extremely important research value in artificial intelligence field.

The ultimate purpose of the dialogue system is to simulate the process of human conversation process and generate human-like responses. Briefly, the dialogue generation problem can be designed as follow: one participator sends a message M, and the agent gives a corresponding response R based on the current message M and the conversation history information C [2].

In the past few decades, dialogue systems draw a great attention in artificial intelligence field. Researchers at many domestic and foreign research institutions and companies have conducted in-depth discussions on related issues. They generally divided dialogue systems into two types on the basis of their functional positioning: task-oriented dialogue systems and nontask-oriented dialogue systems.

The task-oriented dialogue system is also called Closed Domain Dialogue System or Goal Driven Dialogue System, which means that the system has clear service goals or service objects, such as querying restaurants, querying bus lines, querying weather, booking tickets, and ordering meal. In our daily life, DuMi, JIMI, and Siri are all task-oriented dialogue systems. The nontask-oriented dialogue system is also called the Open-Domain Dialogue System. It is mainly based on daily chat, rather than answering specific tasks proposed by users. For example, Microsoft Xiaobing is currently the most famous open-domain dialogue system. This article mainly focuses on Open-Domain Dialogue System.

We review the history of the dialogue system and find that the development of the dialogue system has mainly gone through three stages. At first stage, many dialogue systems are based on rules and frames, that is, the related keywords are set in advance, and a response framework is designed for these keywords. The early rule-based dialogue systems include ELIZA [3], Parry [4], etc.

The retrieval-based dialogue systems [5–7] are the main research direction of the second stage. Since most daily conversations cannot be described by rules or frames, it is difficult for a dialogue system based on rules and frameworks to meet the needs of an open-domain dialogue task. With the great development of the Internet, many resources of human conversations have been accumulated on the social platforms. Since these dialogue resources cover most of the scenarios of conversations, it is possible to obtain candidates through the information retrieval methods and then use the ranking model to select an appropriate response. At the same time, the responses obtained based on the retrieval methods

[*] *Corresponding author. Email: likan@bit.edu.cn*

originate from real human conversations, which are very suitable for the application scenario of the dialogue systems. The research of retrieval dialogue systems basically focuses on the semantic representation, similarity measurement and ranking methods.

At final stage, dialogue systems mainly focus on neural generative conversation models. As one of the main technologies of the dialogue system, generative conversation models can generate a response directly based on the user's message. Compared with the retrieval-based dialogue systems, the structure of the dialogue system based on generative models is relatively simple. Recently, many conversation models are transferred from neural machine translation. In natural language processing (NLP), a neural translation model is a representative task for text generation. It uses deep learning methods to automatically implement text translation, which overcomes the difficulty of constructing generated templates. At the same time, machine translation from one language sentence to another is consistent with the interaction mode in the dialogue. Therefore, it is feasible to build an Open-Domain Dialogue System based on the neural machine translation model.

In this survey, we mainly focus on neural dialogue generation methods in open domain. Section 1 briefly traces the background information and development history of dialogue systems. Section 2 reviews many existing neural dialogue generation methods in open domain. Section 3 summarizes some corpus collections and evaluation indicators of this field. Section 4 discusses some future work of the dialogue system and Section 5 concludes this paper.

2. NEURAL DIALOG GENERATION METHODS

Open-domain dialogue responses generation is an important study in artificial intelligent field. Taking a panoramic view of past approaches, there are six main directions for open-domain dialogue task: *i.e.,* Encoder-Decoder-based methods, Hierarchical Recurrent Encoder-Decoder (HRED)-based methods, Variational AutoEncoder (VAE)-based methods, Reinforcement Learning (RL)-based methods, Generative Adversarial Network (GAN)-based methods and, Pre-training-model-based methods.

In the following subsections, we will detail the basic research of each category, list most existing methods belonging to the category and analyze their advantages, introduce the relationship among different categories, and compare and evaluate the methods of different categories.

2.1. Encoder-Decoder Framework-Based Methods

Sequence-to-sequence (Seq2Seq) model [8,9] builds a great foundation for neural dialogue generation methods. It introduces the Encoder-Decoder framework, and leads to a novel solution for the dialogue generation task. Figure 1 is a general illustration of an Encoder-Decoder framework.

In the case of applying this framework, the encoder encodes the source sequence **X**, a sequence with T tokens concatenated by message and contexts sentences, into a semantic context vector **Context**. Given the **Context** and a response sequence **Y** of length T', the decoder maximizes the generation probability of **Y** conditioned on **Context**: p(**Y**| **Context**).

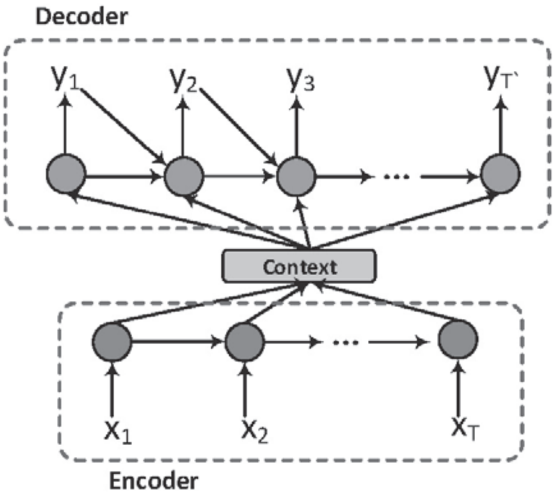

Figure 1 An illustration of the Encoder-Decoder framework.
Source: Chen *et al.* [10].

This Encoder-Decoder framework is the most popular, universal, and basic framework for neural dialogue generation task. Its expansibility is very well, thus resulting in that most existing and novel neural dialogue generation methods are based on this framework. Table 1 shows the summary of these existing methods.

We dive into these existing Encoder-Decoder framework-based methods and divide them into three classifications, which are showed as follows:

1. Adding the external semantic information to control the generating process.

   These methods are simple, which is easy to be thought and implemented. However, these methods also have certain restrictions on the corpus. In addition, some methods also lack the potential for continued research. Meanwhile, the functions for handling the external information are difficult to be changed essentially.

   Persona information: Li *et al.* [11] used persona information to solve the problem of inconsistent response in multi-turn dialogue. However, it only considers the consistency that using different expression but getting the consistent personal information, such as name, address, country, and so on. Moreover, it didn't consider the influence of the dialogue history.

   Cue word information: Yao *et al.* [12] introduced the cue word information to improve the general Seq2Seq model. They proposed a cue word gate recurrent unit to extract the cue word information, and a hierarchical gated fusion unit to fuse this auxiliary information and the general decoding.

   Textual knowledge: Ghazvininejad *et al.* [13] added an encoder to encode the fact message, which could help decoder generate meaningful and proper responses. However, the complexity of this method increases with the increase of knowledge.

Table 1 A summary of existing methods based on Encoder-Decoder framework.

Reference	Dialogue type		Foci of interest in research					Distinguishing characteristics
	Single-turn	Multi-turn	Diversity	Informativeness	Relevance	Consistency	Coherence	
Li *et al.* [11]	✖	✔	✖	✔	✖	✔	✖	Persona information
Yao *et al.* [12]	✔	✖	✖	✖	✔	✔	✔	Cue word gate recurrent unit
Ghazvininejad *et al.* [13]	✔	✖	✔	✔	✖	✖	✖	Fact and knowledge
Huber *et al.* [14]	✔	✖	✔	✔	✔	✖	✖	Emotion information
Tao *et al.* [15]	✔	✖	✔	✔	✔	✖	✖	Constrained multi-head attention
Zhang *et al.* [16]	✔	✖	✔	✔	✔	✖	✖	Specificity control variables
Ko *et al.* [17]	✔	✖	✔	✔	✔	✖	✔	Generating and ordering
Le *et al.* [18]	✔	✖	✔	✔	✔	✖	✖	Multimodal transformer networks (MTN)
See *et al.* [19]	✔	✖	✔	✔	✔	✔	✔	Conditional training, weighted decoding
Cai *et al.* [20]	✖	✔	✔	✔	✖	✖	✔	Group wise, Contrastive learning
He and Glass [21]	✔	✖	✔	✔	✖	✖	✖	Negative training framework
Meditskos *et al.* [22]	✖	✔	✔	✔	✖	✖	✖	Multimodal information, ontology
Su *et al.* [23]	✔	✖	✔	✔	✔	✖	✖	Nonconversational materials

Emotion information: Huber *et al.* [14] introduced emotion information to help dialogue models learn to express emotion when generating a response. They capture the emotion information from the images, including the visual sentiment, facial expression, and scene features. It is the first image-grounded dialogue agent.

Specificity level: Zhang *et al.* [16] proposed the Specificity-based Generation model to deal with the specificity of different utterance-response relations. This module characterized the specificity of the response, which can guide the model to generate responses with different specificities according to different specificity requirements. See *et al.* [19] proposed two controllable neural text generation methods: conditional training and weighted decoding, which controls four important low-level attributes (repetitiveness, specificity, relevance, and question-answer) that affect the quality of a conversation. Those attributes determine whether the response is simple or specific, whether the topic is continues or changes and whether the sentence is a question or answer. Their model achieved the same effect as the giant GPT on some metrics.

Multimodal Information: Le *et al.* [18] proposed the Multimodal Transformer Networks (MTN) to model video information, including image, audio, and text (*e.g.,* subtitles). They employed a single Transformer encoder to encode the text information, and utilized a sliding window of *n* frames and a linear layer to extract video features. Meditskos *et al.* [22] presented a framework for the semantic enrichment and interpretation of communication modalities in dialogue-based interfaces.

2. Using attention mechanism to construct connections between contexts and generated responses.

   The attention mechanism is a good technique for constructing the connections between contexts and responses. It also can show a post hoc analysis for the decoding process. However, the current research on attention mechanism is relatively complete, and the innovative work in this research direction is difficult to appear.

   Normal attention mechanism: Introduced by Bahdanau *et al.* [24] to address the dull responses problem. The idea of attention mechanism is that each token in the response **Y** relays on a different context vector **Context**.

   Self-attention: This method was proposed by Vaswani *et al.* [25] to learn good word vector representations, which is better for natural language understand. Its idea is using other words of the same sentence to recompute the vector representation of one word. This method is widely used in pretraining models.

   Multi-head attention: Tao *et al.* [15] proposed a Constrained Multi-head attention mechanism. They forced the different head attend to different semantics of the same context sentence through a penalty term.

3. Introducing other novel techniques to assist dialogue models.

   These methods always borrow some new research theory and techniques from other field, and change them suitable for dialogue generation task. In general, some methods can easily attract the attention of other researchers. Here are only some novel Encoder-Decoder-based methods that have not yet formed a trend in recent years. As for {HRED, VAE, RL, GAN, and pretraining models}-based methods, we will discuss them in the following sections.

Two-steps generation: Ko *et al.* [17] proposes a Seq2Seq model with attention mechanism. The first part of this model introduced several specificity information during the decoding process to generate responses with different level of specificity. The second component of this model utilized a reordering method based on four external classifiers to increase the semantic rationality of the generated responses. The results of this method are highly depending on the effect of the generation process. Although the ordering process could select rational responses, it doesn't affect the generative capacity.

Novel training framework: He and Glass [21] proposes a new framework named "Negative Training" to address the malicious and frequent responses problem. This framework has two steps: 1) extracting input-output pair exhibits some undesirable behavior (*e.g.,* malicious or frequent responses) and 2) using these pairs as the negative training examples to finetune the model to minimize the model exhibiting bad decoding behavior. They utilized a Seq2Seq model and a training trick of RL structure (*i.e., log derivative trick*) to implement this framework.

Contrastive learning: Recently, the contrastive learning method has attracted much attention in the field of Computer Vision, such as MoCo [26], SimCLR [27] ,and MoCo v2 [28]. Cai *et al.* [20] introduced contrastive learning into dialogue generation, where the model explicitly perceives the difference between the well-chosen positive and negative utterances.

Data enhancement: Su *et al.* [23] selected appropriate responses from nonconversational materials to expand the real corpus collection, which effectively improves the diversity of generated responses. This result of the conversion process has a great influence on the generated response.

2.2. HRED-Based Methods

Since the contexts are not effectively utilized in the general Encoder-Decoder framework, Serban *et al.* [29] introduced the HRED model [30] into dialogue generation task. The difference between Encoder-Decoder framework is that the HRED's encoder consists of two RNNs: one is the token-level RNN and the other one is sentence-level context RNN. A general illustration of encoder in HRED structure is given in Figure 2.

HRED treats a complete dialogue **D** as a sequence of utterances $\{\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_{K-1}, \boldsymbol{U}_K, \boldsymbol{U}_{K+1}\}$. The $\{\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_{K-1}\}$ represents the contexts set, $\boldsymbol{U}_K$ represents the message, and the $\boldsymbol{U}_{K+1}$ represents the response. The token-level RNN model maps each utterance to obtain the final hidden state as the token-level vector. The sentence-level context RNN processes iteratively each token-level vector to understand the semantics of the dialogue. After this processing, the sentence-level context RNN model obtained the hidden states to represent a summary of the dialogue history $(\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_{K-1}, \boldsymbol{U}_K)$, which can rationally handle the longer dialogue history.

HRED is based on Encoder-Decoder framework. The basic HRED is proposed to rationally utilize the dialogue history information to improve the quality of generated responses. Therefore, it is usually used to handle the multi-turn dialogues. In general, most Encoder-Decoder-based methods could replace their framework with the HRED when their targets are changed as multi-turn dialogues. Here we only review some novel methods based on HRED, *e.g.,* WSeq, hierarchical recurrent attention network (HRAN), and ReCoSa. Table 2 shows the summary of these methods.

HRED: The context RNN of HRED encodes the obtained word-level vector which will take advantage of the historical information of the conversation in decoding and generating responses. The purpose of the context RNN is that conducting conversations based on the same conversation background (*e.g.,* topics and concepts), so as to produce meaningful conversations. However, the improvement of HRED over the standard Seq2Seq model is not obvious.

WSeq: Tian *et al.* [31] analyze how to use context effectively through conducting empirical researches to compare various models. Meanwhile, they proposed a variant model named WSeq, which explicitly weights the context vector through context query relevance, and its effect exceeds other benchmark methods.

HRAN: The previous HRED methods pay less attention to the fact that the importance of words and utterances in the context are different. Xing *et al.* [32] proposed a HRAN to attends to important parts within and among utterance-level attention respectively.

ReCoSa: Zhang *et al.* [33] thought a response is always relevant with a few contexts. However, the HRED treats all contexts indiscriminately, which disturbs the generation process. Therefore, they proposed the ReCoSa model based on HRED and self-attention mechanism.

2.3. VAE-Based Methods

In order to generate diverse responses, the VAE [34] was introduced to Encoder-Decoder framework (or HRED). The VAE is a generative model based on a standard autoencoder structure and KL divergence. Figure 3 is a general illustration of VAE model.

The VAE samples the latent variable $\mathscr{Z}$ from a prior probability distribution (*e.g.,* standard Gaussian ($\mu = \vec{0}, \sigma = \vec{1}$)). Then, VAE models learn a posterior recognition model $q(\mathscr{Z}|\boldsymbol{X})$ to replace the deterministic function of autoencoder models. This $q(\mathscr{Z}|\boldsymbol{X})$ is an approximate posterior distribution over $\mathscr{Z}$ (*e.g.,* diagonal Gaussian) conditioned on **X**. Intuitively, the VAE models does not learn the codes as a single point, but as a soft over region in the latent space. When the VAE models encoding the input data into latent variables, rather than taking the input data as independent with each other, they consider the relationships among the input data. Therefore, the latent variables contain certain correlations between each other, which has been proved in Ref. [34]. Therefore, VAE models force the codes to fill the latent space instead of storing the data as an isolated code.

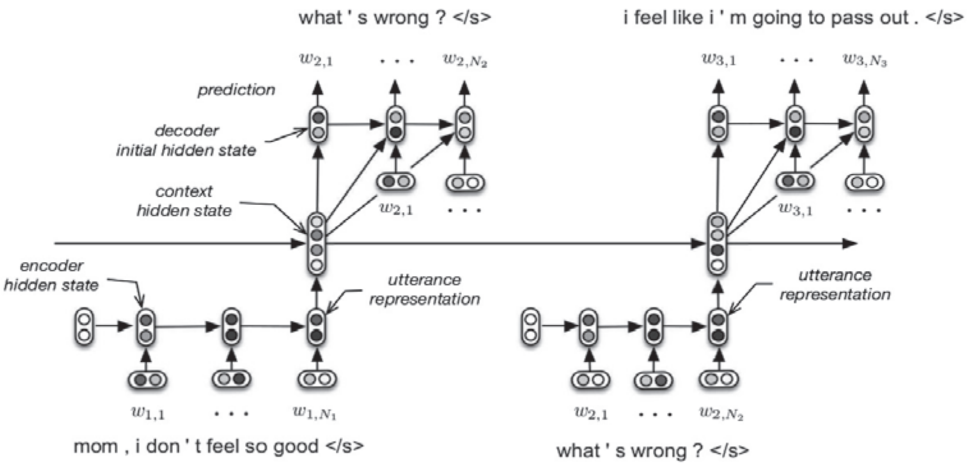
what ' s wrong ? </s>
i feel like i ' m going to pass out . </s>
$w_{2,1}$ . . . w_{2,N_2}
$w_{3,1}$ . . . w_{3,N_3}
prediction
decoder initial hidden state
context hidden state
$w_{2,1}$ . . .
$w_{3,1}$ . . .
encoder hidden state
utterance representation
utterance representation
$w_{1,1}$ . . . w_{1,N_1}
$w_{2,1}$ . . . w_{2,N_2}
mom , i don ' t feel so good </s>
what ' s wrong ? </s>

Figure 2 An illustration of Hierarchical Recurrent Encoder-Decoder (HRED) structure.
Source: Serban *et al.* [29].

Table 2 A summary of existing methods based on Hierarchical Recurrent Encoder-Decoder (HRED).

Reference	Dialogue type		Foci of interest in research					Distinguishing characteristics
	Single-turn	Multi-turn	Diversity	Informativeness	Relevance	Consistency	Coherence	
Serban *et al.* [29]	✖	✔	✔	✔	✔	✖	✖	Hierarchical structure, context-level RNN
Tian *et al.* [31]	✖	✔	✔	✔	✔	✖	✖	Contexts weights
Xing *et al.* [32]	✖	✔	✔	✔	✔	✖	✖	Word-level and utterance-level attention
Zhang *et al.* [33]	✖	✔	✔	✔	✔	✖	✔	HRED model with self-attention

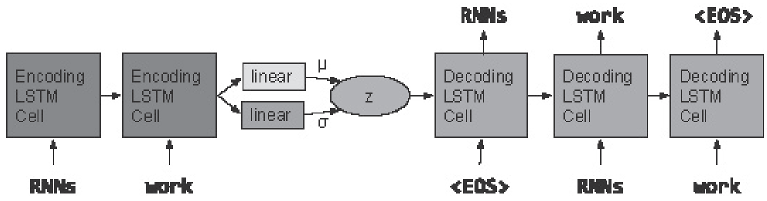

Figure 3 An illustration of the Variational AutoEncoder (VAE) model.
Source: Bowman *et al.* [34].

The VAE introduces latent variables to model the implicit semantic information, and gets well results, which attracts many researches to study the latent variables. The methods based on VAE have achieved good results on diversity metrics. However, according to the Ref. [40], the latent variables may cause incoherent and irrelevant responses. Here we review some dialogue generation methods based on VAE. Table 3 shows the summary of these methods.

VAE+HRED: Serban *et al.* [35] proposed the VHRED model, which is based on the HRED structure and employs a variational module to sample latent variables. They sample the latent variable through the context vector calculated by context-level RNN, which could capture the global semantics. Based on the VHRED model, Chen *et al.* [38] introduced the memory network and proposed VHMN model. They utilized the memory network to record the dialogue history information, and then designed the variational memory reading mechanism to build the context vectors.

Conditional-VAE: Zhao *et al.* [37] proposed a knowledge-guided conditional-VAE (kgCVAE) to utilize dialogue act messages for restraining the latent variables, which improves model effectiveness and interpretability. Meanwhile, they also proposed a new training trick named *bag-of-word-loss* to solve the vanishing latent variable problem [34]. Shen *et al.* [36] also proposed a conditional-VAE(CVAE) model named SPHRED. They designed two status-RNN to encode speaker information and utilize the utterance label to restrict the sampled latent variables. They constructed a classifier to predict the label for the utterance without any labels. Gao *et al.* [39] proposed a discrete CVAE model, which introduces a discrete latent variable with an explicit semantic meaning to improve

Table 3 A summary of existing methods based on Variational AutoEncoder (VAE).

Reference	Dialogue type		Foci of interest in research					Distinguishing characteristics
	Single-turn	Multi-turn	Diversity	Informativeness	Relevance	Consistency	Coherence	
Serban *et al.* [35]	✖	✔	✔	✔	✖	✖	✖	HRED model and latent variables.
Shen *et al.* [36]	✖	✔	✔	✔	✔	✖	✖	Status-RNNs, speaker information
Zhao *et al.* [37]	✖	✔	✔	✔	✔	✖	✖	Knowledge guide, dialog act
Chen *et al.* [38]	✖	✔	✔	✔	✔	✖	✔	Memory network, VAE and HRED
Gao *et al.* [39]	✔	✖	✔	✔	✔	✖	✖	Two-stage approach discrete CVAE
Gao *et al.* [40]	✔	✔	✔	✔	✔	✖	✔	Joint optimization, spacefusion model

the general CVAE on dialogue generation task. They proposed a two-stage sampling approach to enable efficient diverse variable selection from a large latent space assumed in the dialogue generation task.

SPACEFUSION: In order to address the irrelevant responses problem caused by random sampled latent variables, Gao *et al.* [40] proposed a joint optimized model named SPACEFUSION. They utilized multi-task training framework to train a Seq2Seq model and an autoencoder, and then designed an interpolation term to implement the fusion process of the two latent space of Seq2Seq and autoencoder.

2.4. RL-Based Methods

Also, in order to generate diverse responses on multi-turn dialogue generation task, the RL method chose another solution. The RL method focuses on the optimization process. Li *et al.* [41] thought the Maximum Likelihood Estimation (MLE) is sensitive to high-frequency sentences, thus resulting in safe responses with less information. Therefore, they introduced RL algorithm and designed reward functions to replace the MLE process.

RL algorithm is widely used in goal-oriented dialogue task [42–45], which shows the potential for improving response quality in open-domain dialogue generation task. Here we only review some RL grounded methods on open-domain dialogue generation task. Table 4 shows the summary of these methods.

RL: Li *et al.* [41] introduced RL into dialogue generation task. They focused on the design of the reward function. They illustrated that "ease of answering," "information flow," and "semantic coherence" are three main factors that could promote the success of a dialog. Then, they proposed the approximate reward functions to model the three factors.

PRGDDA: Yang *et al.* [46] proposed a RL grounded method named Personalized Response Generation by Dual-learning-based Domain Adaptation (PRGDDA). They first trained a generative model through a large dataset without persona information, and then fine-tuned the model with a small size personalized data by using dual-learning mechanism.

Seq2SeqCo-{bi, MP, dual}: Zhang *et al.* [47] thought the reason that Seq2Seq always generates the dull responses is the optimization function equals the Kullback–Leibler divergence. Therefore, they replaced the original optimization function with the coherence score. They proposed three models *i.e.*, GRU Bilinear (bi), MatchPyramid (MP), and Dual-Learning Architecture (dual), to calculate the three coherence scores, respectively.

Multiple-Response-Generation-Model: Gao *et al.* [48] proposed a response generation model to generate multiple diverse responses simultaneously. Their model considered a set of responses jointly, and contained a latent word inference network to sample a discrete word that related with the context and response. They utilize the RL algorithm to optimize their model.

P^2 BOT: Based on the RL algorithm, Liu *et al.* [49] introduced mutual persona perception and proposed a transmitter-receiver framework to explicitly model the interaction between participators of one conversation. This method focuses on the personalized dialogue generation.

2.5. GAN-Based Methods

Since the GAN could not effectively handle the discrete sequence [50], it is hard to obtain a good result through GAN for dialogue generation task. To address this problem, Yu *et al.* [50] and Li *et al.* [51] introduced policy gradient method, and Xu *et al.* [52] proposed an approximate embedding layer to help GAN handle the discrete situation.

After this, many novel GAN based generation methods have been proposed, such as MaskGAN [53], DP-GAN [54], Adver-REGS [51], GAN-AEL [52], Adversarial Information Maximization (AIM) [55], DialogWAE [56], and Posterior-GAN [57]. Here we only review some GAN grounded methods for dialogue generation task. Table 5 shows the summary of these methods.

Adver-REGS: To improve the performance of GAN on dialogue generation task, Li *et al.* [51] employed the policy gradient algorithm of RL method. They proposed rewards for every generation step (REGS) to solve the disadvantage that the expectation of rewards is approximated through only one sample, and the reward is used for all actions. They proposed two strategies to compute each

Table 4 A summary of existing methods based on Reinforcement Learning (RL).

Reference	Dialogue type		Foci of interest in research					Distinguishing characteristics
	Single-turn	Multi-turn	Diversity	Informativeness	Relevance	Consistency	Coherence	
Li *et al.* [41]	✔	✔	✔	✖	✔	✖	✔	Ease of answering, information flow, semantic coherence
Yang *et al.* [46]	✔	✔	✔	✔	✔	✔	✖	Dual-learning, domain adaptation
Zhang *et al.* [47]	✔	✖	✔	✔	✔	✖	✔	Coherence reward, dual-learning archi-tecture
Gao *et al.* [48]	✔	✖	✔	✔	✔	✖	✖	Latent word inference network
Liu *et al.* [49]	✖	✔	✔	✔	✔	✔	✔	Mutual persona perception, personalized dialog

Table 5 A summary of existing methods based on Generative Adversarial Network (GAN).

Reference	Dialogue type		Foci of interest in research					Distinguishing characteristics
	Single-turn	Multi-turn	Diversity	Informativeness	Relevance	Consistency	Coherence	
Li *et al.* [51]	✔	✔	✔	✔	✔	✖	✔	Reward for every generation step
Xu *et al.* [52]	✔	✖	✔	✔	✔	✖	✖	Approximate embedding layer
Zhang *et al.* [55]	✔	✖	✔	✔	✔	✖	✖	Adversarial information maximization
Gu *et al.* [56]	✖	✔	✔	✔	✔	✖	✔	Wasserstein distance, VAE idea
Feng *et al.* [57]	✖	✔	✔	✔	✔	✖	✔	Future information, query-response-future

step reward: (1) Monte Carlo search; (2) training discriminator to assign rewards for partially decoding sentences.

GAN-AEL: Xu *et al.* [52] proposed an approximate embedding layer to replace the sample process, which makes the adversarial training process become a derivable process. This method alleviates the instability problem when using RL training algorithm to a certain extent.

AIM: Zhang *et al.* [55] proposed an embedding-based structured discriminator and developed AIM model to generate informative and diverse responses.

Dialog-WAE: Gu *et al.* [56] used GANs to train potential distributions. It used a neural network to generate context-dependent random "noise" that is sampled from the prior and posterior distributions of potential variables, and minimized the Wasserstein distance between the two distributions. Then, a Gaussian mixture prior network is used to enrich the latent space.

Posterior-GAN: Feng *et al.* [57] proposed a novel posterior adversarial learning framework to utilize the future dialogue information. They reconstructed the original multi-turn dialogue dataset, *i.e.*, replacing the original query-response pairs to the query-response-future triples. They also proposed two Encoder-Decoder-based discriminators (*i.e.*, a forward discriminator and a backward discriminator) to cooperatively discriminate the coherence and informativeness of the generated response through query and future information, respectively.

2.6. Pretraining-Model-Based Methods

In recent years, the pretraining models have a huge impact in the field of natural language understanding and natural language generation. Mehri *et al.* [58] studied the sentence representation based on the pretraining models and employed the pretrained representation in dialogue generation task. They proved that the pretraining model can be utilized in open-dialogue generation task. Some works employ pretraining models to construct dialogue systems, such as DialoGPT [59], Blender [60], Meena [61], and Plato-2 [62]. However, due to the huge cost of training a pretraining model for open-domain dialogue generation task, it is not suitable for the individual researchers. Table 6 shows the summary of these methods.

DialoGPT: Zhang *et al.* [59] proposed a large tunable dialogue model named DialoGPT that based on the GPT-2 model [63]. They also introduced the Maximum Mutual Information (MMI) to address the dull responses problem.

Table 6 A summary of existing methods based on pretraining models.

Reference	Dialogue type		Foci of interest in research					Distinguishing characteristics
	Single-turn	Multi-turn	Diversity	Informativeness	Relevance	Consistency	Coherence	
Zhang *et al.* [59]	✖	✔	✔	✔	✔	✔	✔	GPT-2 structure, mutual information maximization
Roller *et al.* [60]	✖	✔	✔	✔	✔	✔	✔	Blended skill talk
Adiwardan *et al.* [61]	✖	✔	✔	✔	✔	✔	✔	Evolved transformer, seq2seq model
Bao *et al.* [62]	✖	✔	✔	✔	✔	✔	✔	Unified network, curriculum learning, latent variable.

Blender: Roller *et al.* [60] proposed a large-scale dialogue model named Blender. To further analyze the effectiveness of their methods, they build variants of this model with 90M, 2.7B, and 9.4B parameters. They introduced the Blended Skill Talk (BST) to help Blender learn the conversation skills: (1) providing fascinating viewpoints; (2) listening carefully to their partner; (3) demonstrating knowledge, empathy, and personality at the right time; (4) keeping their personality consistent.

Meena: Adiwardana *et al.* [61] proposed an open-domain chatbot named Meena that has a single evolved transformer [64] encoder and 13 Evolved Transformer decoder. They train the best model for 30 days on a TPUv3 Pod (2,048 TPU cores) on the Meena dataset containing 40B words (or 61B BPE tokens).

PLATO-2: Bao *et al.* [62] proposed a dialogue model named PLATO-2. They introduced curriculum learning framework to training the latent variable of PLATO-2. The PLATO-2 employed the unified network architecture and contained two parts models in Chinese and English. The Chinese model was trained on the 1.2B Chinese open-domain multi-turn dialogue corpus, while the English model was trained on the 700M English open-domain multi-turn corpus. They train their model for 3 weeks on 64 Nvidia V100.

2.7. Comparison of Different Categories of Methods

Many neural dialogue generation methods have been proposed in recent years. Taking a panoramic view of these approaches, we roughly divide them into six categories that has shown above. However, their basic researches have strong relationship between each other. The earliest Encoder-Decoder framework-based dialogue generation method was migrated from machine translation field and widely called as Seq2Seq model. Due to the limitation of the Seq2Seq on handling dialogue history, the HRED model was proposed. Since both Seq2Seq and HRED were difficult to consider the latent information that hiding in the dialogue, The VAE model was proposed, which generated diverse responses through sampled latent variables. The RL employed two Seq2Seq-agents to simulate the dialogue, and designed the reward functions to achieve the optimize these two agents. Therefore, the agent of RL often generates purposeful and consistent responses, which is more suitable for task-oriented dialogue generation task. The application of RL pushed the GAN into the open-domain dialogue generation task. Aiming at the problem that the traditional GAN is hardly trained by the discrete output, Li *et al.* [51] introduced the policy gradient method, which effectively address the problem. As for the pretraining models, it can be traced to the transformer model, which is also based on the Encoder-Decoder framework.

Based on these basic models of each category, many novel and effective methods are proposed in recent years. At present, many studies currently focus on different motivations and uses different datasets and different evaluation metrics, thus resulting in the difficulty of doing specific comparison results. Therefore, we evaluated and analyzed the results of each method and gave a simple comparison under as objective conditions as possible.

In general, the Seq2Seq without the attention mechanism, a simple conversation model based on Encoder-Decoder framework, is the worst-performing neural generative model. It often generates “safe reply” which lacking diversity, informativeness, relevance, consistency, and coherence. On the opposite, the pre-training-model-based methods can get the-state-of-the-art performance in general dialogue generation task. However, it is not proper for individual researchers because it costs too much. Besides the pretraining-model-based methods, the five remaining categories of methods are all research hotspots for most researchers of NLP field.

During the past few years, researchers always focused on the diversity, informativeness, relevance, consistency, and coherence of the generated responses, and proposed many novel methods. Since the basic Seq2Seq model often generates safe and dull responses, some researchers introduced the attention mechanism to learn the semantic relationships between contexts and responses, some researchers utilized the external semantic information to assistant in generating dialogue responses, and some researchers proposed the HRED model to handle the dialogue history when generating responses. HRED, the foundational method of the HRED-based methods, adds sentence-level RNN to the Seq2Seq model, and is mainly used for processing the multi-turn dialogues. Since HRED can extract the context vector from the dialogue history, the responses will be generated to make the dialogue continue. However, experiments show that the HRED does not significantly improve the evaluation results, and it is only slightly better than the Seq2Seq with attention.

After a short while, some researchers introduced the latent variables to model the hidden information of dialogue because a good response often not only related to the dialogue history but also to the hidden information that out of the dialogue. The CVAE, a foundational method of VAE-based dialogue generation methods, significantly increases the diversity and informativeness of generated responses through introducing the latent variable sampled from a Gaussian distribution. However, the sampled latent variables also cause incoherent and irrelevant responses. Therefore, the subsequent works are mainly focusing on improving the relevance and coherence of responses while maintaining the good diversity and informativeness.

With the development of the theories and techniques of deep learning field, some new methods could be transferred to dialogue generation task (*e.g.,* RL method). Since the RL is often used for achieving some goals, it is suitable for multi-turn dialogue generation task, which needs purposed and consistent responses. Therefore, the RL-based methods can easily generate coherent responses. Moreover, the RL-based methods can design multiple reward functions for different targets. However, training the RL-based methods is difficult, so some works were proposed to increase the stability of training the RL-based methods.

Another famous deep learning technique, the GAN, also draws much attentions for dialogue generation task. The GAN-based methods reduce safe and general responses and increase the diversity, informativeness, relevance, consistency, and coherence of responses through changing optimization algorithm (*e.g.,* AdverREGS [51]) or designing novel discriminators (*e.g.* DP-GAN [54]). Whereas, same with the RL-based methods, the GAN-based methods are also difficult to train.

In summary, for these five foci of interesting in research, the GAN-based methods generally have the best performance among the remaining five categories. The performance of RL-based methods are slightly weak than GAN-based methods, but better than the VAE-based methods in general. The performance of VAE-based methods reach a middle stage of the remaining five categories. Although the diversity and informativeness of VAE-based methods are much the same as GAN-based and RL-based methods in sometimes, the relevance and coherence of VAE-based methods are weak than both RL-based and GAN-based methods. The performance of HRED-based methods are generally weak than GAN, RL, and VAE-based methods, but better than Seq2Seq models.

Although the Seq2Seq model generally has the worst performance, its research value is not small. In recent two years, many researchers proposed novel methods to improve the Seq2Seq model and enriched the Encoder-Decoder framework-based methods on other interesting targets. For example, See *et al.* [19] focused on controlling the responses generation through low-level attributes. He and Glass [21] focused on addressing the malicious and frequent responses problem, Su *et al.* [23] focused on enhancing the dialogue dataset. These methods generally have the state-of-the-art results on their targets.

3. DIALOG DATASETS AND EVALUATION METRICS

This section reviews some open-domain dialog datasets and existing evaluation metrics.

3.1. Dialog Datasets

The dialogue corpus promotes the development of the automatic dialogue system. With the development of the Internet, the form of communication between people has gradually shifted from simple face-to-face to major Internet social platforms, such as Sina-Weibo, Douban, Facebook, Twitter, etc. This transformation has allowed conversations to be stored on social platforms in the form of text, voice, and even video. The single-turn and multi-turn dialogue resources accumulated in social platforms provide quantities of corpus for the research of dialogue systems, and also make the construction of automatic dialogue systems feasible. What is more critical is that the dialogue rules and modes contained in real dialogue resources will promote the research of dialogue systems. Table 7 shows several dialogue corpuses, including OpenSubtitles,[1] CornellMovie,[2] sina-weibo, Ubuntu,[3] DailyDialog,[4] DoubanConversationCorpus,[5] PersonaChat,[6] and STC-SeFun. In addition, we have also sorted out two multi-modal data sets, namely MUStARD[7] and CH-SIMS.[8] Although they are not standard dialog data, they can be processed into dialog data for dialog generation.

3.2. Evaluation Metrics

The basic technology of the generative dialogue system originates from the task of machine translation. Therefore, the evaluation metrics of the generative dialogue systems also inherit the evaluation metrics of machine translation field such as BLEU [76]. In addition, the relevance between responses and user's messages is an important issue of dialogue systems, so some relevance evaluation indicators between texts (such as the word vector-based relevance measurement method) are also introduced into the dialogue evaluation process. However, the evaluation object of these indicators is a single round of dialog, and the dialogue is a continuous multi-round process, so the quality of a single round dialogue cannot reflect the overall performance of the dialogue system, especially for the purpose of communication. Table 8 shows some general auto-evaluation metrics.

Generally speaking, the Perplexity metric is used for evaluating the convergent degree of the dialogue systems. Bleu and Embedding-based metrics can reflect the Relevance of the response and the context to a certain extent. Distinct metric is widely utilized to represent the Diversity of the generated responses. Coherence is the metric to evaluate the coherence of response and context. As for the

[1] http://opus.nlpl.eu/OpenSubtitles-v2018.php

[2] http://www.cs.cornell.edu/~cristian/Cornell Movie-Dialogs Corpus.html

[3] https://github.com/rkadlec/ubuntu-ranking-dataset-creator

[4] http://yanran.li/dailydialog

[5] https://github.com/MarkWuNLP/MultiTurnResponseSelection

[6] https://github.com/facebookresearch/ParlAI/tree/master/parlai/tasks/personachat

[7] https://github.com/soujanyaporia/MUStARD

[8] https://github.com/thuiar/MMSA

Table 7 | A summary of open-domain dialog datasets.

No	Name	Description	Year	Ref.
1	Open Subtitles	The dataset is constructed from subtitles of a large number of movies (one segmentation every 3 sentences, and the third sentence as a reply).	2009	[65]
2	Cornell Movie	The dataset contains of fictional conversations extracted from raw movie scripts.	2011	[66]
3	Sina-weibo	The dataset of short-text conversation from Sina Weibo (Chinese microblog). This dataset provides rich collection of instances and consists of both natural dialogs, human-generated labels, and lots of candidate responses.	2013	[67]
4	Ubuntu	The dataset consists of over 7 million sentences, which formed almost 1 million multi-turn dialogs. The dataset has both the multi-turn property of dialogs and the unstructured nature of the interaction from microblog.	2015	[68]
5	Daily-Dialog	The high-quality multi-turn dialogue dataset. The dialogues in this dataset reflect human's daily communication way and involve various topics about human's daily life.	2017	[69]
6	Douban Conversaion Corpus	The dataset construct from douban.com. In this dataset, there are 10 responses as candidates for each context. Each candidate has three labels to judge if it is proper for the session. A proper response can naturally reply to the given context.	2017	[70]
7	Persona Chat	The dataset was collected by Amazon MechanicalTurk. It contains 162,064 dialog sentences from humans, with a maximum of 15 words per sentence for each sentence. The humans are randomly paired, and each person is randomly assigned a personalized role.	2018	[71]
8	STC-SeFun	The dataset consists of short-text conversation pairs with their sentence functions manually annotated.	2019	[72]
9	MUStARD	The dataset contains a total of 690 videos with a total duration of about 9626 seconds. The data source is American comedy on youtube. The dataset contains high-quality artificial annotations, such as satire expression, context, speaker, video, and audio tags.	2019	[73]
10	CH-SIMS	The dataset contains 2,281 refined video segments in the wild with both multi-modal and independent unimodal annotations.	2020	[74]

Informativeness and Consistency are usually evaluated by human evaluation.

Recently, the research of automatic evaluation of open-domain dialogue generation draws much attention. Pang *et al.* [80] proposed that using the GPT-2 model as the standard to automatically measure the quality of the generated responses, including context coherency, response fluency and diversity, and logical self-consistency. Mehri and Eskenazi [81] proposed an unsupervised automatic evaluation method with less references. They used RoBERTa to automatically measure the quality of the generated responses, and found the results have a high correlation with the effect of human evaluation.

4. FUTURE OUTLOOK OF NEURAL GENERATIVE DIALOGUE SYSTEMS

This paper mainly researches and analyzes the structures and technologies of the existing nontask generative dialogue systems to some extent, and introduces some open datasets and evaluation metrics. We believe that the research work in this field can be improved or opened up new research directions from the following entry points.

1. The introduction of knowledge will improve the performance of the dialogue systems. In general, a good conversation always involves many aspects of knowledge (*e.g.*, background knowledge, personal information, emotional information, etc.). In the real-world conversations, participators often give proper responses based on their own knowledge. The knowledge is on one hand the basis for understanding the conversation, and on the other hand is the key point to facilitate the dialogue. In the past few years, many researchers utilized the knowledge to control the generation and improved the quality of the generated responses, which shows the potential of the knowledge. However, it yet reaches the stage of fully and effectively using knowledge. Therefore, how to effectively utilize the knowledge is still a key problem because this information is really important for the conversation models.

2. Multidisciplinary (*e.g.*, aesthetics, psychological, sociology) theories and methods can be introduced into the neural dialogue generation methods for increasing the performance. Generally speaking, a person's aesthetics, psychological activities, social status, and other factors will affect his (or her) external expression in a conversation. Most factors are studied and concluded as theories and methods, which can be transferred to the dialogue generation task. Although most of neural dialogue generation methods are data-driven and do not need to consider the details, it is a feasible research route to fuse existing conversation models with classical theories and methods in other disciplines to improve the quality of the generated responses. Therefore, with the development of theories

Table 8 A summary of auto-evaluation metrics of dialog systems.

No	Metric Name	Description	Year	Ref.
1	Perplexity	The metric calculated using probability, similar to information entropy. Generally, when a language model is used to evaluate the probability of a test sentence, the higher the probability, the lower the perplexity, and the better the language model.	1992	[75]
2	BLEU	The word-overlapping-based metric, which calculates word overlapping degree between the generated response and the ground-truth response.	2002	[76]
3	Distinct	The widely used metric calculates the percentage (%) of distinct n-gram, which reflects the degree of diversity of the generated responses.	2016	[77]
4	Embedding based	Including embedding average, embedding greedy, and embedding extrema, embedding-based metrics first calculate semantic embedding based on the vectors of all individual tokens in responses and then calculate the similarity between the generated response and the real-world response by cosine distance.	2016	[78]
5	Coherence	The metric refers to the coherence of responses and contexts. It is the averaged word embedding similarity between the words of the context and the response computed using embedding vectors	2018	[79]

and methods in other disciplines, there will be some progress in the open field dialogue generation method.

3. The combination of the pretraining models and the existing effective general methods may bring more state-of-the-art results. At present, the methods based on pre-training models often reach the good performance through training on a very large-scale dataset with a large cost of time and hardware resources. However, their model structure has not been changed much, *e.g.*, Meena [61] used Evolution Transformer to replace the general Transformer, PLATO2 introduced latent variables to model implicit semantic information. It is possible to improve generation results through integrating the existing novel and effective methods with the pretraining-model-based methods. However, due to the large amount of time and hardware resources, this research direction is not suitable for individual researchers.

4. For dialogue generation task in open domain, the automatic evaluation metrics are not consistent to human ratings to some extent. At present, the evaluation of the dialogue system is mainly realized through automatic evaluation metrics and manual evaluation. Due to the high cost and low efficiency of manual evaluation, it is difficult to quickly evaluate the conversation models and improve the research speed. The traditional automatic evaluation metrics mainly come from the machine translation field, and to a certain extent it is difficult to meet the needs of evaluating the dialogue systems. The new automatic evaluation metrics based on the pretraining model that appeared in the past two years are difficult to convince everyone because the similarity between them and the human evaluation results is still not high. It is a good research direction to propose some new automatic evaluation metrics that are highly consistent with human evaluation.

5. The open datasets of the dialogue system are difficult to coordinate in academic research and dialogue applications. Good-quality datasets need to be constructed artificially, with low efficiency, and may be different from actual application scenarios after construction, which is suitable for research but not suitable for practical applications. The data extracted from the actual dialogue scene, such as Reddit, Twitter, movie subtitles, etc., have some problems in terms of quality, quantity, and the uncertain semantics. Large amounts of data but poor quality is not conducive to academic research. Since the datasets of dialogue are different, the conversation context cannot be effectively used for every model, which is trained by one or two datasets. The quality of dialogue systems is also based on the datasets.

5. CONCLUSION

In this survey, we summarize the neural dialogue generation methods in open domain. The construction method is still mainly based on the retrieval methods and the generative methods. These two kinds of methods have their own advantages and disadvantages. We focus on the open-domain dialogue systems and review the main generative dialogue methods in this survey. The generative dialogue methods can directly generate a response based on the semantics of the context and the user messages, without being restricted by the dialogue resource library. However, as a key research direction in the field of NLP, there are still many issues that need to be further explored. The current generation of conversational technology is still immature that results in dull and general responses. The relations between automatic evaluation metrics and human ratings are less well understand. Finally, we enumerated some feasible research directions for the neural dialogue generation task in open domain.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORS' CONTRIBUTIONS

Bin Sun shaped the framework of this paper and provided sufficient information to the development of the main content. Bin Sun and Kan Li contributed to the writing of the manuscript and Kan Li has provided critical feedback and suggestive support to this work. All authors contributed to the final manuscript.

Funding Statement

This research is supported by National Key R&D Program of China (No. 2016YFB0801100), Beijing Natural Science Foundation (No. 4172054, L181010), and National Basic Research Program of China (No. 2013CB329605). Kan Li is the corresponding author.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their valuable and constructional advices on the previous versions of this article; all remaining errors are our own.

REFERENCES

[1] A.M. Turing, Computing machinery and intelligence, Mind. 49 (1950), 433–460.

[2] T. Fong, C.E. Thorpe, C. Baur, Collaboration, dialogue, human-robot interaction, in: R.A. Jarvis, A. Zelinsky (Eds.), ISRR, Springer Tracts in Advanced Robotics, vol. 6, Springer, Berlin, Heidelberg, Germany, 2001, pp. 255–266.

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

[4] K.M. Colby, F.D. Hilf, S. Weber, H.C. Kraemer, Turing-like indistinguishability tests for the calidation of a computer simulation of paranoid processes, Artif. Intell. 3 (1972), 199–221.

[5] D. Ameixa, L. Coheur, P. Fialho, P. Quaresma, Luke, I am your father: dealing with out-of-domain re-quests by using movies subtitles, in: T. Bickmore, S. Marsella, C. Sidner (Eds.), Intelligent Virtual Agents, IVA, Lecture Notes in Computer Science, vol. 8637, Springer, Cham, Switzerland, 2014, pp. 13–21.

[6] W.H. Gomaa, A.A. Fahmy, A survey of text similarity approaches, IJCA. 68 (2014), 13–18.

[7] X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W.X. Zhao, *et al.*, Multi-turn response selection for chatbots with deep attention matching network, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018, pp. 1118–1127.

[8] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in Twenty-eighth Conference on Neural Information Processing Systems (NeurIPS-2014), Palais des Congrès de Montréal, Montréal CANADA, 2014, pp. 3104–3112. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

[9] L. Shang, Z. Lu, H. Li, Neural responding machine for short-text conversation, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 2015, pp.1577–1586.

[10] H. Chen, X. Liu, D. Yin, J. Tang, A survey on dialogue systems: recent advances and new frontiers, SIGKDD Explor. 19 (2017), 25–35.

[11] J. Li, M. Galley, C. Brockett, G.P. Spithourakis, J. Gao, W.B. Dolan, A persona-based neural conversation model, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 2016.

[12] L. Yao, Y. Zhang, Y. Feng, D. Zhao, R. Yan, Towards implicit content-introducing for generative short-text conversation systems, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017, pp. 2190–2199.

[13] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J Gao, W. Yih, *et al.*, A knowledge-grounded neural conversation model, the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USAA, 2018, pp. 5110–5117. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16710

[14] B. Huber, D.J. McDuff, C. Brockett, M. Galley, B. Dolan, Emotional dialogue generation using image-grounded language models, in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, Canada, 2018, 277.

[15] C. Tao, S. Gao, M. Shang, W. Wu, D. Zhao, R. Yan, Get the point of my utterance! Learning towards effective responses with multi-head attention mechanism, in the 27th International Joint Conference on Artificial Intelligence (IJCAI-2018) and the 23rd European Conference on Artificial Intelligence, the premier international gathering of researchers in AI! IJCAI-ECAI-18 is part of the Federated AI Meeting (FAIM), Stockholm, Sweden, 2018, pp. 4418–4424.

[16] R. Zhang, J. Guo, Y. Fan, Y. Lan, J. Xu, X. Cheng, Learning to control the specificity in neural response generation, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018, pp. 1108–1117.

[17] W. Ko, G. Durrett, J.J. Li, Linguistically-informed specificity and semantic plausibility for dialogue generation, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2019, pp. 3456–3466.

[18] H. Le, D. Sahoo, N.F. Chen, S.C.H. Hoi, Multimodal transformer networks for end-to-end video-grounded dialogue systems, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 5612–5623.

[19] A. See, S. Roller, D. Kiela, J. Weston, What makes a good conversation? How controllable attributes affect human judgments, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2019, pp. 1702–1723.

[20] H. Cai, H. Chen, Y. Song, Z. Ding, Y. Bao, W. Yan, *et al.*, Group-wise contrastive learning for neural dialogue generation, in Findings of the Association for Computational Linguistics (EMNLP 2020), 2020, pp. 793–802.

[21] T. He, J.R. Glass, Negative training for neural dialogue response generation, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2044–2058.

[22] G. Meditskos, E. Kontopoulos, S. Vrochidis, I. Kompatsiaris. Converness: ontology-driven conversational awareness and context understanding in multimodal dialogue systems, Expert Syst. J. Knowl. Eng. 37 (2020), e12378.

[23] H. Su, X. Shen, S. Zhao, X. Zhou, P. Hu, R. Zhong, *et al.*, Diversifying dialogue generation with non-conversational text, in

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7087–7097.

[24] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, ICLR, in 3th International Conference on Learning Representations, (ICLR-2015), The Hilton San Diego Resort & Spa, 2015. http://arxiv.org/abs/1409.0473

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Attention is all you need, in NIPS, Thirty-first Conference on Neural Information Processing Systems (NeurIPS-2017), Long Beach Convention Center, Long Beach, 2017, pp. 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[26] K. He, H. Fan, Y. Wu, S. Xie, R.B. Girshick, Momentum contrast for unsupervised visual representation learning, CoRR abs/1911.05722, 2019. http://arxiv.org/abs/1911.05722

[27] T. Chen, S. Kornblith, M. Norouzi, G.E. Hinton, A simple framework for contrastive learning of visual representations, CoRR abs/2002.05709, 2020. https://arxiv.org/abs/2002.05709

[28] X. Chen, H. Fan, R.B. Girshick, K. He, Improved baselines with momentum contrastive learning, CoRR abs/2003.04297, 2020. https://arxiv.org/abs/2003.04297

[29] I.V. Serban, A. Sordoni, Y. Bengio, A.C. Courville, J. Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models, in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI), Phoenix Convention Center, Phoenix, Arizona, USA, 2016, pp. 3776–3784. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957

[30] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J.G. Simonsen, J. Nie, A hierarchical recurrent encoder- decoder for generative context-aware query suggestion, in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM), Melbourne, Australia, 2015, pp. 553–562.

[31] Z. Tian, R. Yan, L. Mou, Y. Song, Y. Feng, D. Zhao, 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 (Volume 2: Short Papers), Vancouver, Canada, 2017, pp. 231–236.

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

[33] H. Zhang, Y. Lan, L. Pang, J. Guo, X. Cheng, 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, Florence, Italy, 2019, pp. 3721–3730.

[34] S.R. Bowman, L. Vilnis, O. Vinyals, A.M. Dai, R. Józefowicz, S. Bengio, Generating sentences from a continuous space, in Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), Berlin, Germany, 2016, pp. 10–21.

[35] I.V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A.C. Courville, *et al.*, A hierarchical latent variable encoder-decoder model for generating dialogues, in the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-2017), San Francisco, California, USA, 2017, pp. 3295–3301. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567

[36] X. Shen, H. Su, Y. Li, W. Li, S. Niu, Y. Zhao, *et al.*, A conditional variational frame- work for dialog generation, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada, 2017, pp. 504–509.

[37] T. Zhao, R. Zhao, M. Eskenazi, Learning discourse-level diversity for neural dialog models using conditional variational autoencoders, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, 2017, pp. 654–664.

[38] H. Chen, Z. Ren, J. Tang, Y.E. Zhao, D. Yin, Hierarchical variational memory network for dialogue generation, in Proceedings of the 2018 World Wide Web Conference (WWW), Lyon, France, 2018, pp. 1653–1662.

[39] J. Gao, W. Bi, X. Liu, J. Li, G. Zhou, S. Shi, A discrete CVAE for response generation on short-text conversation, 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), Hong Kong, China, 2019, pp. 1898–1908.

[40] X. Gao, S. Lee, Y. Zhang, C. Brockett, M. Galley, J. Gao, *et al.*, Jointly optimizing diversity and relevance in neural response generation, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 1229–1238.

[41] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, J. Gao, Deep reinforcement learning for dialogue generation, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 2016, pp. 1192–1202.

[42] P. Su, M. Gasic, N. Mrksic, L.M. Rojas-Barahona, S. Ultes, D. Vandyke, *et al.*, Continuously learning neural dialogue management, CoRR abs/1606.02689, 2016. http://arxiv.org/abs/1606.02689

[43] T. Wen, D. Vandyke, N. Mrksic, M. Gasic, L.M. Rojas-Barahona, P. Su, *et al.*, A network-based end-to-end trainable task-oriented dialogue system, in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (EACL), Valencia, Spain, 2017, pp. 438–449.

[44] A. Das, S. Kottur, J.M.F. Moura, S. Lee, D. Batra, Learning cooperative visual dialog agents with deep reinforcement learning, in IEEE International Conference on Computer Vision (ICCV-2017). ICCV, Venice, Italy 2017, pp. 2970–2979.

[45] M. Lewis, D. Yarats, Y.N. Dauphin, D. Parikh, D. Batra, Deal or no deal? End-to-end learning for negotiation dialogues, CoRR abs/1706.05125, 2017.

[46] M. Yang, W. Tu, Q. Qu, Z. Zhao, X. Chen, J. Zhu, Personalized response generation by dual-learning based domain adaptation, Neural Netw. 103 (2018), 72–82.

[47] H. Zhang, Y. Lan, J. Guo, J. Xu, X. Cheng, Reinforcing coherence for sequence to sequence model in dialogue generation, in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 2018, pp. 4567–4573.

[48] J. Gao, W. Bi, X. Liu, J. Li, S. Shi, Generating multiple diverse responses for short-text conversation, in Proceedings of the AAAI

Conference on Artificial Intelligence (AAAI), Hilton Hawaiian Village, Honolulu, Hawaii, USA, 2019, pp. 6383–6390.

[49] Q. Liu, Y. Chen, B. Chen, J. Lou, Z. Chen, B. Zhou, *et al.*, You impress me: dialogue generation via mutual persona perception, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1417–1427.

[50] L. Yu, W. Zhang, J. Wang, Y. Yu, SeqGAN: sequence generative adversarial nets with policy gradient, in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI), Hilton San Francisco, San Francisco, California, USA, 2017, pp. 2852–2858. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14344

[51] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, D. Jurafsky, Adversarial learning for neural dialogue generation, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017, pp. 2157–2169.

[52] Z. Xu, B. Liu, B. Wang, C. Sun, X. Wang, Z. Wang, *et al.*, Neural response generation via GAN with an approximate embedding layer, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017, pp. 617–626.

[53] W. Fedus, I.J. Goodfellow, A.M. Dai, MaskGAN: better text generation via filling in the ____, in ICLR (Poster) 6th International Conference on Learning Representations, (ICLR-2018), Vancouver, BC, Canada, 2018. https://openreview.net/forum?id=ByOExmWAb

[54] J. Xu, X. Ren, J. Lin, X Sun, 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 (EMNLP), Brussels, Belgium, 2018, pp. 3940–3949.

[55] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, Generating informative and diverse conversational responses via adversarial information maximization, in 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, 2018, pp. 1815–1825. https://proceedings.neurips.cc/paper/2018/hash/23ce1851341ec1fa9e0c259de10bf87c-Abstract.html

[56] X. Gu, K. Cho, J. Ha, S. Kim, DialogWAE: multimodal response generation with conditional Wasserstein auto-encoder, in 7th International Conference on Learning Representations (ICLR-2019), New Orleans, LA, USA, 2019. https://openreview.net/forum?id=BkgBvsC9FQ

[57] S. Feng, H. Chen, K. Li, D. Yin, Posterior-GAN: towards informative and coherent response generation with posterior generative adversarial network, in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Hilton New York Midtown, New York, New York, USA, 2020, pp. 7708–7715.

[58] S. Mehri, E. Razumovskaia, T. Zhao, M. Eskenazi, Pretraining methods for dialog context representation learning, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 3836–3845.

[59] Y. Zhang, S. Sun, M. Galley, Y. Chen, C. Brockett, X. Gao, *et al.*, DIALOGPT: large-scale generative pre-training for conversational response generation, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 270–278.

[60] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, *et al.*, Recipes for building an open-domain chatbot, CoRR abs/2004.13637, 2020. https://arxiv.org/abs/2004.13637

[61] D. Adiwardana, M. Luong, D.R. So, J. Hall, N. Fiedel, R. Thoppilan, *et al.*, Towards a human-like open-domain chatbot, CoRR abs/2001.09977, 2020. https://arxiv.org/abs/2001.09977

[62] S. Bao, H. He, F. Wang, H. Wu, H. Wang, W. Wu, *et al.*, PLATO-2: towards building an open-domain Chatbot via curriculum learning, CoRR abs/2006.16779, 2020. https://arxiv.org/abs/2006.16779

[63] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, Technical Report, OpenAI, 2019. https://www.openai.com/blog/better-language-models/

[64] D.R. So, Q.V. Le, C. Liang, The evolved transformer, in Proceedings of the 36th International Conference on Machine Learning (ICML, 2019), Long Beach Convention & Entertainment Center in Long Beach, California, 2019, pp. 5877–5886. http://proceedings.mlr.press/v97/so19a.html

[65] J. Tiedemann, News from OPUS—a Collection of Multilingual Parallel Corpora with Tools and Interfaces, Recent Advances in Natural Language Processing, 5 (2009), pp. 237–248.

[66] C. Danescu-Niculescu-Mizil, L. Lee, Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs, in Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, Portland, OR, USA, 2011. https://www.aclweb.org/anthology/W11-0609/

[67] H. Wang, Z. Lu, H. Li, E. Chen, A dataset for research on short-text conversations, in the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP-2013), Grand Hyatt Seattle, Seattle, Washington, USA, 2013, pp. 935– 945. https://www.aclweb.org/anthology/D13-1096/

[68] R. Lowe, N. Pow, I. Serban, J. Pineau, The Ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems, in Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czech Republic, 2015, pp. 285–294.

[69] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, S. Niu, DailyDialog: a manually labelled multi-turn dialogue dataset, in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan, 2017, pp. 986–995. https://www.aclweb.org/anthology/I17-1099/

[70] Y. Wu, W. Wu, C. Xing, M. Zhou, Z. Li, Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots, in 55th Annual Meeting of the Association for Computational Linguistics (ACL-2017), Vancouver, Canada, 2017, pp. 496–505.

[71] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, Personalizing dialogue agents: I have a dog, do you have pets too?, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018, pp. 2204–2213.

[72] W. Bi, J. Gao, X. Liu, S. Shi, Fine-grained sentence functions for short-text conversation, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 3984–3993.

[73] S. Castro, D. Hazarika, V. Perez-Rosas, R. Zimmermann, R. Mihalcea, S. Poria, Towards multimodal SarcasmDetection (an obviously PerfectPaper), in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 4619–4629.

[74] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, *et al.*, CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained

annotation of modality, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3718–3727.

[75] P.F. Brown, S.D. Pietra, V.J.D. Pietra, J.C. Lai, R.L. Mercer, An estimate of an upper bound for the entropy of English, Comput. Linguistics, 18 (1992), 31–40.

[76] K. Papineni, S. Roukos, T. Ward, W Zhu, Bleu: a method for automatic evaluation of machine translation, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Grenoble, France, 2002, pp. 311–318.

[77] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A diversity-promoting objective function for neural conversation models, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 2016, pp. 110–119.

[78] C. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, J. Pineau, How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 2016, pp. 2122–2132.

[79] X. Xu, O. Dusek, I. Konstas, V. Rieser, Better conversations by modeling, filtering, and optimizing for coherence and diversity, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 2018, pp. 3981–3991.

[80] B. Pang, E. Nijkamp, W. Han, L. Zhou, Y. Liu, K. Tu, Towards holistic and automatic evaluation of open-domain dialogue generation, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3619–3629.

[81] S. Mehri, M. Eskenazi, USR: an unsupervised and reference free evaluation metric for dialog generation, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 681–707.