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

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 [3–5, 7] are the main research direction of the second stage. 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...
originates 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 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 $Context: p(Y|Context)$.

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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dialogue type</th>
<th>Foci of interest in research</th>
<th>Distinguishing characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li et al. [11]</td>
<td>Single-turn</td>
<td>✗</td>
<td>Person information</td>
</tr>
<tr>
<td></td>
<td>Multi-turn</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diversity</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relevance</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coherence</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Yao et al. [12]</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Ghazvininejad et al. [13]</td>
<td></td>
<td>✔</td>
<td>Fact and knowledge</td>
</tr>
<tr>
<td>Huber et al. [14]</td>
<td></td>
<td>✗</td>
<td>Emotion information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td>Constrained multi-head attention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Tao et al. [15]</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Zhang et al. [16]</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Ko et al. [17]</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Le et al. [18]</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>See et al. [19]</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Cai et al. [20]</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>He and Glass [21]</td>
<td></td>
<td>✗</td>
<td>Negative training framework</td>
</tr>
<tr>
<td>Meditskos et al. [22]</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Su et al. [23]</td>
<td></td>
<td>✗</td>
<td>Nonconversational materials</td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✗</td>
<td></td>
</tr>
</tbody>
</table>

**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 \( \text{Context} \).

**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 fine-tune 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 \{U_1, U_2, ..., U_{K-1}, U_K, U_{K+1}\}. The \{U_1, U_2, ..., U_{K-1}\} represents the contexts set, \(U_K\) represents the message, and the \(U_{K+1}\) 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 \(\{U_1, U_2, ..., U_{K-1}, U_K\}\), 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 \(\mathcal{Z}\) from a prior probability distribution (e.g., standard Gaussian (\(\mu = 0, \sigma = 1\))). Then, VAE models learn a posterior recognition model \(q(\mathcal{Z}|X)\) to replace the deterministic function of autoencoder models. This \(q(\mathcal{Z}|X)\) is an approximate posterior distribution over \(\mathcal{Z}\) (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.
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 2** A summary of existing methods based on Hierarchical Recurrent Encoder-Decoder (HRED).

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

---

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

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

2.5. GAN-Based Methods

Since the GAN could not effectively handle the discrete sequence generation task, 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
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.

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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dialogue type</th>
<th>Foci of interest in research</th>
<th>Distinguishing characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Diversity</td>
<td>Informativeness</td>
</tr>
<tr>
<td>Li et al. [41]</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Yang et al. [46]</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Zhang et al. [47]</td>
<td>✔</td>
<td>✖</td>
<td>✔</td>
</tr>
<tr>
<td>Gao et al. [48]</td>
<td>✔</td>
<td>✖</td>
<td>✔</td>
</tr>
<tr>
<td>Liu et al. [49]</td>
<td>✖</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dialogue type</th>
<th>Foci of interest in research</th>
<th>Distinguishing characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Diversity</td>
<td>Informativeness</td>
</tr>
<tr>
<td>Li et al. [51]</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Xu et al. [52]</td>
<td>✔</td>
<td>✖</td>
<td>✔</td>
</tr>
<tr>
<td>Zhang et al. [55]</td>
<td>✔</td>
<td>✖</td>
<td>✔</td>
</tr>
<tr>
<td>Gu et al. [56]</td>
<td>✖</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Feng et al. [57]</td>
<td>✖</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Table 6  A summary of existing methods based on pretraining models.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dialogue type</th>
<th>Foci of interest in research</th>
<th>Distinguishing characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-turn</td>
<td>Multi-turn</td>
<td>Diversity</td>
</tr>
<tr>
<td>Zhang et al. [59]</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Roller et al. [60]</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Adiwardana et al. [61]</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Bao et al. [62]</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

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 assist 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., Adversary-based metrics can reflect the relevance between responses and user’s messages is not standard dialog data, they can be processed into dialog data for evaluation process. However, the evaluation object of these metrics 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 7 shows several dialogue corpuses, including OpenSubtitles, CornellMovie, sina-weibo, Ubuntu, DailyDialog, DoubanConversationCorpus, PersonaChat, and STC-SeFun. In addition, we have also sorted out two multi-modal data sets, namely MUSTARD and CH-SIMS. Although they are not standard dialog data, they can be processed into dialog data for dialog generation.

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, CornellMovie, sina-weibo, Ubuntu, DailyDialog, DoubanConversationCorpus, PersonaChat, and STC-SeFun. In addition, we have also sorted out two multi-modal data sets, namely MUSTARD and CH-SIMS. 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

2 http://www.cs.cornell.edu/~cristian/Cornell Movie-Dialogs Corpus.html
3 https://github.com/rikadlec/ubuntu-ranking-dataset-creator
4 http://yanran.li/dailyydialog
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/MMISA

3. DIALOG DATASETS AND EVALUATION METRICS

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

---

1. **OpenSubtitles**: A large collection of subtitles from movies, TV shows, and other media.
2. **Cornell Movie**: A dataset for movie reviews.
3. **sina-weibo**: A dataset for social media interactions.
4. **Ubuntu**: A dataset for technical support conversations.
5. **DailyDialog**: A dataset for daily conversational dialogue.
6. **DoubanConversationCorpus**: A dataset for conversations from Douban.
7. **PersonaChat**: A dataset for persona-based conversations.
8. **STC-SeFun**: A dataset for storytelling and conversation.
10. **CH-SIMS**: A dataset for conversational human-machine interaction.

---

**References**


A summary of open-domain dialog datasets.

<table>
<thead>
<tr>
<th>No</th>
<th>Name</th>
<th>Description</th>
<th>Year</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Open Subtitles</td>
<td>The dataset is constructed from subtitles of a large number of movies (one segmentation every 3 sentences, and the third sentence as a reply).</td>
<td>2009</td>
<td>[65]</td>
</tr>
<tr>
<td>2</td>
<td>Cornell Movie</td>
<td>The dataset contains fictional conversations extracted from raw movie scripts.</td>
<td>2011</td>
<td>[66]</td>
</tr>
<tr>
<td>3</td>
<td>Sina-weibo</td>
<td>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.</td>
<td>2013</td>
<td>[67]</td>
</tr>
<tr>
<td>4</td>
<td>Ubuntu</td>
<td>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.</td>
<td>2015</td>
<td>[68]</td>
</tr>
<tr>
<td>5</td>
<td>Daily-Dialog</td>
<td>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.</td>
<td>2017</td>
<td>[69]</td>
</tr>
<tr>
<td>6</td>
<td>Douban Conversation Corpus</td>
<td>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.</td>
<td>2017</td>
<td>[70]</td>
</tr>
<tr>
<td>7</td>
<td>Persona Chat</td>
<td>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.</td>
<td>2018</td>
<td>[71]</td>
</tr>
<tr>
<td>8</td>
<td>STC-SeFun</td>
<td>The dataset consists of short-text conversation pairs with their sentence functions manually annotated.</td>
<td>2019</td>
<td>[72]</td>
</tr>
<tr>
<td>9</td>
<td>MUSTARD</td>
<td>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.</td>
<td>2019</td>
<td>[73]</td>
</tr>
<tr>
<td>10</td>
<td>CH-SIMS</td>
<td>The dataset contains 2,281 refined video segments in the wild with both multi-modal and independent unimodal annotations.</td>
<td>2020</td>
<td>[74]</td>
</tr>
</tbody>
</table>

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, participants 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
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 pretrained 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 constructive advices on the previous versions of this article; all remaining errors are our own.

REFERENCES

[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.


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


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


