Challenges in Building Intelligent Open-domain Dialog Systems

MINLIE HUANG and XIAOYAN ZHU, Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
JIANFENG GAO, Microsoft Research, WA, USA

There is a resurgent interest in developing intelligent open-domain dialog systems due to the availability of large amounts of conversational data and the recent progress on neural approaches to conversational AI [33]. Unlike traditional task-oriented bots, an open-domain dialog system aims to establish long-term connections with users by satisfying the human need for communication, affection, and social belonging. This paper reviews the recent work on neural approaches that are devoted to addressing three challenges in developing such systems: *semantics*, *consistency*, and *interactiveness*. *Semantics* requires a dialog system to not only understand the content of the dialog but also identify user's emotional and social needs during the conversation. *Consistency* requires the system to demonstrate a consistent personality to win users trust and gain their long-term confidence. *Interactiveness* refers to the system's ability to generate interpersonal responses to achieve particular social goals such as entertainment and conforming. The studies we select to present in this survey is based on our unique views and are by no means complete. Nevertheless, we hope that the discussion will inspire new research in developing more intelligent open-domain dialog systems.

CCS Concepts: • **Information systems** → *Information systems applications*; *Users and interactive retrieval*; • **Computing methodologies** → **Natural language processing**; **Machine learning**; **Discourse, dialogue and pragmatics**; **Natural language generation**; **Neural networks**.

Additional Key Words and Phrases: dialog system, chatbot, social bot, conversation generation, response generation, conversational AI

ACM Reference Format:
Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in Building Intelligent Open-domain Dialog Systems. *ACM Transactions on Information Systems* 1, 1, Article 1 (January 2020), 33 pages. https://doi.org/10.1145/3383123

1 INTRODUCTION

Building intelligent open-domain dialog systems that can converse with humans coherently and engagingly has been a long-standing goal of artificial intelligence (AI). Early dialog systems such as Eliza [151], Parry [18], and Alice [142], despite being instrumental to significantly advancing machine intelligence, worked well only in constrained environments. An open-domain social bot remains an elusive goal until recently. The Microsoft XiaoIce ('Little Ice' literally in Chinese) system, since its release in May, 2014, has attracted millions of users and can converse with users on a wide

Authors' addresses: Minlie Huang; Xiaoyan Zhu, Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, Beijing, China, aihuang@tsinghua.edu.cn; Jianfeng Gao, Microsoft Research , Redmond, WA, USA, jfgao@microsoft.com.

1046-8188/2020/1-ART1 $15.00
https://doi.org/10.1145/3383123

variety of topics for hours [126, 192]. In 2016, the Alexa Prize challenge was proposed to advance the research and development of social bots that are able to converse coherently and engagingly with humans on popular topics such as sports, politics, and entertainment, for at least 20 minutes [16, 107] [1] . The evaluation metric, inspired by the Turing Test [138], is designed to test the social bots' capacity of delivering coherent, relevant, interesting, free-form conversations and keeping users engaged as long as possible. However, the general intelligence demonstrated by these systems is still far behind humans. Building open-domain dialog systems that can converse on various topics like humans remains extremely challenging [33].

In this paper we focus our discussion on three challenges in developing neural-based open-domain dialog systems, namely *semantics*, *consistency* and *interactiveness*. The rest of the paper is structured as follows. In the rest of Section 1, we compare open-domain dialog bots with traditional task-oriented bots and elaborate the three challenges. In Section 2, we survey three typical approaches to building neural-based open-domain dialog systems, namely, retrieval-based, generation-based, and hybrid methods. In Sections 3, 4, and 5, we review the approaches that have been proposed to address the three challenges, respectively. In Section 6, we discuss recent work on open-domain dialog evaluation. In Section 7, we present an incomplete survey of frequently-used or recently-proposed benchmarks for open-domain conversation modeling. We conclude the paper by presenting several future research trends in Section 8.

1.1 Open-Domain Dialog vs. Task-Oriented Dialog

Generally speaking, there are two types of dialog systems: task-oriented and open-domain dialog. Task-oriented dialog systems are designed for specific domains or tasks, such as flight booking, hotel reservation, customer service, and technical support, and have been successfully applied in some real-world applications. Open-domain dialog systems, however, are much more challenging to develop due to its open-ended goal.

As outlined by Gao et al. [33], although both task-oriented dialog and open-domain dialog can be formulated as an optimal decision making process with the goal of maximizing expected reward, the reward in the former is better-defined and much easier to optimize than the latter. Consider a ticket-booking bot. It is straightforward to optimize the bot to get all necessary information to have the ticket booked in minimal dialog turns. The goal of an open-domain dialog agent is to maximize the long-term user engagement. This is difficult to optimize mathematically because there are many different ways (known as dialog *skills*) to improve the engagement (e.g., providing entertainment, giving recommendations, chatting on an interesting topic, providing emotional comforting) and it requires the systems to have a deep understanding of dialog context and user's emotional needs to select the right skill at the right time, and generate interpersonal responses with a consistent personality.

Open-domain dialog systems also differ from task-oriented bots in system architecture. A task-oriented bot is typically developed based on a pre-defined task-specific schema[2] and is designed as a modular system which consists of domain-specific components like language understanding, dialog management[3], and language generation[4]. These components can be either hand-crafted based on domain knowledge or trained on task-specific labeled data. On the other hand, due to

[1]Even though the dialog systems in this challenge are very complicated, they are more informational systems where user emotion need is less considered.
[2]A task schema typically defines a set of user intents, and for each intent defines a set of dialog acts, slot-value pairs.
[3]Dialog management performs both dialog state tracking [47, 89] and response selection via policy [71, 99, 132, 183].
[4]Recently, there are end-to-end methods [9, 112, 180] that output a response given the previous dialog history. But in general, domain knowledge about the task needs to be explicitly considered, which differs significantly from open-domain dialog systems.

the open-ended nature, open-domain dialog systems need to deal with open-domain knowledge without any pre-defined task-specific schemas or labels. In recent years, there has been a trend towards developing fully data-driven, end-to-end systems that map user's input to system's response using neural networks. Since the primary goal of open-domain dialog bots is to be AI companions to humans with an emotional connection rather than completing specific tasks, they are often developed to mimic human conversations by training neural response generation models on large amounts of [123, 129, 141].

Unlike task-oriented bots, most neural response generation models developed for open-domain dialog systems are not grounded in real world, which prevents these systems from effectively conversing about anything that relates to the user's environment. Only recently have researchers begun to explore how to ground open-domain dialog systems in real-world entities and knowledge [36, 88, 104]. Knowledge grounding is also crucial for the system to provide interpersonal responses. For instance, the conversations between friends are quite different from those between strangers. So the system needs to be grounded in the personas of the speaker and addressee, respectively [62]. The tone of system responses needs to be adjusted according to user's emotional states and affect by grounding in affect or emotion of the user [52, 154, 162].

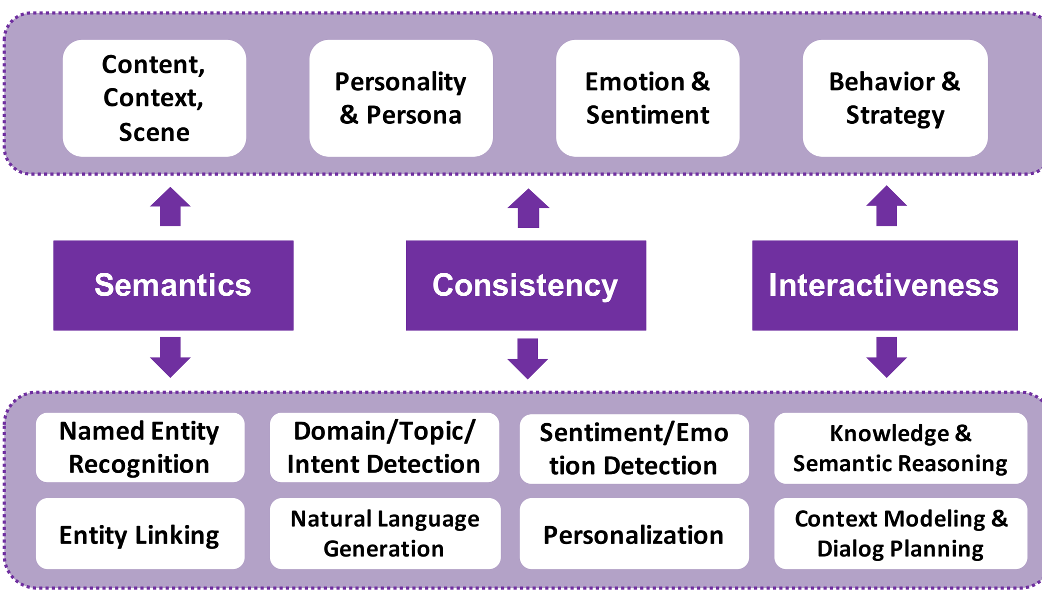

Fig. 1. Key issues (the middle block), key factors (the upper block), and key technologies (the bottom block) in open-domain dialog systems.

1.2 Three Challenges

Thanks to the recent success of neural approaches to natural language processing and conversational AI, there have been substantial advancements on open-domain dialog systems.

Most neural approaches formulate conversation as a task of generating output response given user input and dialog context. Almost all neural response generation models follow the neural text generation frameworks, such as sequence-to-sequence (Seq2Seq) [123, 133, 141], conditional variational autoencoder (CVAE) [127], and generative adversarial network (GAN) [64]. These models are trained on large-scale conversational data, and can deliver plausible responses to any topic that users are interested in. However, these neural-based systems often fail to establish long-term emotional connections with users due to many issues as illustrated by the examples in Table 1 and summarized in Figure 1. In what follows, we elaborate three key issues which we will discuss in the rest of the paper [5].

Semantics Issue	
User	I come from Shanghai.
Machine	Cool! I am from Beijing, we come from the same city!
User	Back to Beijing now. It is so cold but I am very happy!
Machine	When will you come to Beijing?
Consistency Issue	
User	Are a boy or a girl?
Machine	I am a boy.
User	Are you a girl?
Machine	Yes, I am a girl.
Interactiveness Issue	
User	Terrible breakfast today.
Machine	How about together tomorrow morning?
User	Sure. Would be great with your recommendation.
Machine	Really?

Table 1. Examples for demonstrating the semantics, consistency, and interactiveness issues. The first two examples show semantic conflicts between user post and machine response, the third session shows bad consistency within the dialog context due to the lack of a coherent personality, and the last session has bad interactiveness due to the lack of grounding. The results in the first two blocks are from a standard Seq2Seq model with an attention mechanism, and the last session is from a commercial system.

Semantics. Semantics is the heart of any dialog system because conversation is a semantic activity [190]. It requires not only to *understand the content of the conversation*, which is often in multi-modality including text, image and video, but also to *understand users* by identifying information beyond the dialog content such as a user's personality and persona[6], emotion, sentiment, and the user's profile and background. From the technical perspective, semantics mainly involves the key techniques of *natural language understanding and user understanding*, including named entity recognition, entity linking, domain detection, topic and intent detection, user sentiment/emotion/opinion detection, and knowledge/ commonsense reasoning.

Consistency. In order to gain user's long-term confidence and trust, it is crucial for a dialog system to present consistent behaviors and respond consistently given user's input and dialog history [62, 103, 186, 192]. For instance, a social bot should not deliver a response that conflicts with her pre-set persona, or her previous responses in temporal dependency, causality, or logic. Specifically, the system's response needs to be consistent in three dimensions. First is persona consistency where the response needs to fit the pre-defined personality of the dialog system. Second is stylistic consistency where a consistent speaking style is presented. Third is contextual consistency in which the response needs to be coherent and consistent with respect to the dialog context. From the technical perspective, consistency mainly involves personalization, stylistic generation, and multi-turn context modeling.

[5]Note that the challenges discussed in this section are also fundamental to traditional, non-neural dialog systems.
[6]Personality is someone's character or nature while a persona is a superficial identity of the character or nature.

Interactiveness. As mentioned above, meeting user's social needs, such as emotional affection and social belonging, is the primary design goal of an open-domain dialog system. Interactiveness refers to *the system's ability to achieve complex social goals such as entertainment and conforming by optimizing its behaviors and dialog strategies in multi-turn conversation.* To improve interactiveness, it is important to understand the user's emotion state or affect [189, 192], to respond not only reactively but also proactively [108, 148, 169], to control the topic maintenance or transition [146], and to optimize the interaction strategy (i.e., dialog policy) in multi-turn conversations to maximize long-term user engagement. From the technical perspective, interactiveness mainly involves sentiment and emotion detection, dialog state tracking, topic detection and recommendation, dialog policy learning, and controllable response generation.

We summarize the techniques required to address the three issues in Figure 1, including named entity recognition, entity linking, domain/topic/intent detection, and sentiment/emotion detection. As demonstrated in the Alexa Prize challenge which targets at developing dialog systems for conversing coherently and engagingly with humans on various popular topics, the winning dialog systems [16, 29] are composed of different modules that are developed based on these techniques, including language understanding, dialog management, and natural language generation. In such modular designs, the semantic issue is mainly related to the understanding module which is intended to understand the dialog (e.g., content, entity, topic, etc.) and user (e.g., opinion, personality, emotional needs). The other two issues are mainly related to the dialog management and generation modules, aiming to generate responses that are not only consistent in content and personality, but also interactive so as to increase the long-term user engagement. These issues are highly interleaved. For example, understanding dialog and user (semantics) is fundamental to generating consistent and interactive responses.

2 FRAMEWORKS FOR BUILDING OPEN-DOMAIN DIALOG SYSTEMS

As discussed in Section 1.1, open-domain dialog systems are typically implemented using an end-to-end architecture, rather than a modular architecture used by task-oriented bots for which task-specific schemas and labels are available for the development of these dialog modules. At the heart of an open-domain dialog system is a response generation engine, which takes user input at t-th dialog turn $X_t = x_1^t x_2^t \cdots x_n^t$ and dialog context C_t, which will be explained in a minute, and generates response $Y_t = y_1^t y_2^t \cdots y_m^t$ as

$$\hat{Y}_t = \arg\max_{Y \in \Omega} \mathcal{P}_\theta(Y|X_t, C_t) \tag{1}$$

where Ω denotes the set of all candidate responses, $\mathcal{P}_\theta$ is a learned model of scoring candidate responses, parameterized by θ, and argmax the search algorithm to find among all candidates the best one with the highest score.

This formulation unifies three typical methods of building open-domain dialog systems: retrieval-based, generation-based, and hybrid. In retrieval-based methods, the search space Ω is obtained by retrieving candidate responses from a pre-collected human conversational dataset consisting of input-context-response pairs. $\mathcal{P}_\theta(Y|X_t, C_t)$ is implemented as a matching or ranking function which scores the relevance of each candidate given X_t and C_t. In generation-based methods, the search space Ω is very large, namely $Y \in V^m$ where V is the vocabulary size and m is the response length, and $\mathcal{P}_\theta(Y|X_t, C_t)$ is typically implemented as an auto-regressive model that generates a sentence word by word. In the hybrid methods, it is typical to first retrieve *prototype* responses from a dataset and then generates a response by utilizing prototype responses.

Note that the introduction of context C_t offers a lot of flexibility to model various aspects of dialog. For instance, when $C_t = \emptyset$, it models single-turn dialog; Setting $C_t = X_1 Y_1 X_2 Y_2 \cdots X_{t-1}$ models multi-turn dialogs. C_t can also encode other (non-content) contexts such as persona [103, 175, 186] for personalized dialog generation, emotion labels [4, 189] for emotional response generation, and knowledge graphs [36, 190] for knowledge-aware response generation.

2.1 Retrieval-based Methods

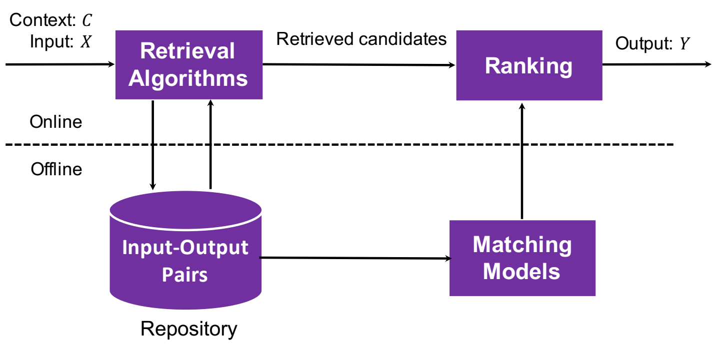

Fig. 2. Framework of retrieval-based methods. The online process finds the most relevant output from the retrieved candidate with a matching model while the offline process trains the matching model with the auto-constructed data.

Given a dialog corpus and the user's post, IR-based systems can use any retrieval algorithm to choose an appropriate response from the corpus [13, 53, 60]. In such a setting, the system retrieves the most similar post to the given user post, and the response to the retrieved post is returned as the response to the user's post. Traditional learning-to-rank methods were introduced by Ji et al. [54] for response selection from a large-scale post-response repository. Afterwards, many neural models have been proposed. Figure 2 illustrates the process of retrieval-based response generation methods. Using input $X \oplus C$ [7] as a query, such methods first retrieve a list of candidates from a large repository which consists of input-context-output pairs, and choose the top-scored candidate as output response Y using the matching function $\mathcal{P}_\theta(Y|X, C)$, which can be implemented using either traditional learning-to-rank algorithms [75], or modern neural matching models [28, 51, 80]. The model parameters θ is commonly learned by minimizing the margin-based pair-wise ranking loss as follows[8]:

$$\mathcal{L} = max(0, \gamma + match_\theta(Y_-, X \oplus C) - match_\theta(Y_+, X \oplus C)) \tag{2}$$

where γ is a margin (a hyper-parameter), Y_+ is a ground-truth (positive) response, Y_- is a negative response which can be randomly sampled from the dataset or generated by corrupting Y_+, and $match_\theta(Y, X \oplus C)$ is the matching function to be learned.

Alternatively, we can also use a likelihood loss defined as:

$$\begin{gathered}\mathcal{L} = -\log \mathcal{P}_\theta(Y_+|X \oplus C) \\ \mathcal{P}(Y_+|X \oplus C) = \frac{\exp\{match_\theta(Y_+, X \oplus C)\}}{\exp\{match_\theta(Y_+, X \oplus C)\} + \sum_{i=1}^{k} \exp\{match_\theta(Y_-^i, X \oplus C)\}}\end{gathered} \tag{3}$$

[7]Hereafter, we will use $X \oplus C$ to denote the input query that combines the current user input X and the dialog context C.
[8] Note that the method of pair-wise ranking is widely used in the literature, but other ways such as point-wise and list-wise ranking methods [75] are also feasible.

Modern neural models of $match(Y, X \oplus C)$ can be roughly grouped into two categories, shallow and deep interaction networks[9], as illustrated in Figure 3. In shallow interaction networks, candidate Y and input $X \oplus C$ are first encoded independently into the two vectors which then have some *shallow* interactions such as subtraction or element-wise multiplication before being fed to the classification layer. In deep interaction networks, Y and $X \oplus C$ interact via an interaction network to form a fused representation, which is then fed to the classification layer.

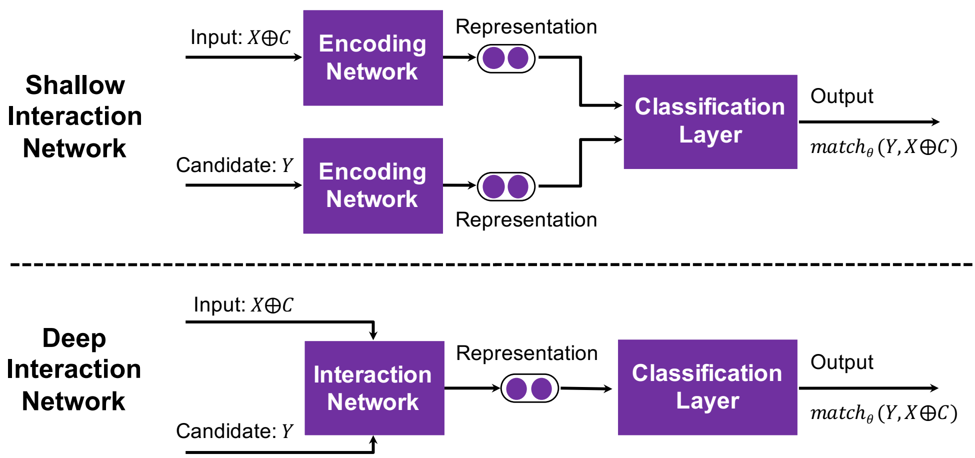

Fig. 3. Frameworks of shallow and deep interaction networks. In shallow interaction network, the feature vectors of input $X \oplus C$ and candidate Y are obtained independently, and there may be shallow interactions such as subtraction or element-wise multiplication between the two vectors before the classification layer. In deep interaction network, the input and candidate make interactions in the early stage to obtain a feature vector for the classification layer.

For shallow interaction networks, many efforts have been devoted to learning good representations for query and candidate independently. Huang et al. [51] proposed to use deep structured similarity models (DSSMs) to extract semantic features from query and document independently before computing their relevance. DSSM is further augmented by introducing Convolutional layers [34, 49, 122, 125] and recurrent layers with Long Short-Term Memory (LSTM) units [95]. To effectively incorporate dialog history, Yan et al. [163] reformulated input query X, and combined matching scores computed based on the reformulated and original queries, and retrieved queries and responses, respectively. Zhou et al. [193] used a hierarchical Recurrent Neural Network (RNN) to encode a candidate and the utterance sequence in context, respectively, before computing their matching score. These shallow models are simple to implement and efficient to execute.

For deep interaction networks, query $X \oplus C$ and response Y interact via a neural network to generate a single feature vector that preserves all query-response interaction information at different levels of abstraction. The matching score is then derived from the vector using another neural network. Hu et al. [49] extracted matching features from all n-gram combinations of input X and response Y to obtain low-level feature maps with a Convolutional Neural Network (CNN). Afterwards, the feature maps are transformed with multiple CNN layers to form the final representation for classification. Wu et al. [159] proposed a sequential matching network (SMN) for multi-turn

[9]Shallow or deep is regarding *interaction*, namely whether the learned representations are obtained by early-stage interactions (deep), or late-stage (sometimes no) interactions (shallow). The two words are not referring to whether the model structure is deep or not.

dialog where each contextual utterance in $X \oplus C$ is encoded conditioned on Y, and these utterances are connected sequentially by GRUs. The matching score is computed on top of the weighted sum of the GRUs' states. Zhou et al. [194] proposed a deep attention matching network. The query and its candidate response are firstly represented with self-attention inspired by the transformer network [139], and then the interactions between them were made with cross-attention to obtain word-by-word matching matrices, and finally the matching score is computed by aggregating all the matching information with a 3D matching tensor. Yang et al. [166] extended SMN with external knowledge in information-seeking conversation systems. The method first expands response candidates using pseudo-relevance feedback, and then makes the candidates interact with the query to obtain word-by-word matching matrices. The subsequent operations are very similar to SMN. Zhang et al. [181] proposed a deep utterance aggregation model which shares a similar structure with SMN. The difference lies in that gated self-attention was used to obtain the representations of the query and a response candidate, and the subsequent operations are almost the same to SMN. Wu et al. [157] proposed to consider topic clues for query-response matching. The authors first extracted topical words for the query and response respectively using LDA. Then, a query representation is conditioned not only on the response representation but also on the attentive read of the topical words of the response. A response representation is computed similarly conditioned on the message's topical words and the query representation. Other matching models that were proposed originally for non-dialog tasks such as paraphrase detection, language inference, and reading comprehension [97, 149], have also been adapted and applied to dialog response ranking.

One of the most notable deep interaction networks for learning the matching function (as defined by Eq. 2) is BERT [23], which achieves state-of-the-art performance on many NLP tasks, including response selection. $X_t \oplus C_t$ and a candidate response y, normally separated by a special token [SEP], form the input of a multi-layer Transformer [139] blocks (12-48 blocks). Each block consists of multi-head a self-attention module, layer normalization, a feed forward layer, and residual connections. The vectors at the output layer are fed to a fine-tuned classifier to determine whether the response y is appropriate for the input. This structure has been widely adopted in retrieval-based methods [46].

There is a short review on deep retrieval-based dialogue systems [10] where the authors discussed existing work with respect to single-turn matching models, multi-turn matching models, and ensemble models. In comparison, we summarize existing work from the interaction perspective: whether a candidate response makes deep matching with the input (post, or along with the context) at early or late stage. In general, deep interaction networks usually work better than shallow interaction networks [137].

2.2 Generation-based Methods

Neural generative models have been widely applied to open-domain dialog generation. Inspired by the early template-based generation method [48] and statistical machine translation (SMT) [111], sequence-to-sequence (Seq2seq) models [123, 129, 133, 141] have become the most popular choice for dialog generation. Other frameworks, including conditional variational autoencoder (CVAE) [26, 56, 121, 124, 184, 185] and generative adversarial network (GAN) [64, 161], are also applied to dialog generation. Very recently, Transformer-based language models pretrained with large-scale corpora are another popular choice [39, 106, 155, 179], which obtains strong performance in dialog generation [155].

Generation-based models usually formulate $\mathcal{P}(Y|X_t \oplus C_t)$ as:

$$\mathcal{P}(Y|X_t \oplus C_t) = \prod_{i=1}^{m} P(y_i|y_{<i}; X_t \oplus C_t). \tag{4}$$

where $y_{<i} = y_1 y_2 \cdots y_{i-1}$. Typically, the output response is generated word by word, e.g., at each time step a word is sampled according to $P(y|y_{<i}; X_t \oplus C_t)$. Using RNNs, during the course of generation, the generated prefix is autoregressively encoded into the input to generate the next word.

Most neural generation models adopt an encoder-decoder framework. The encoder transforms the input $X_t \oplus C_t$ into semantic vectors as

$$\mathbf{X}_t \oplus \mathbf{C}_t = \mathbf{Encoder}(X_t \oplus C_t). \tag{5}$$

Then, at each i-th step of generation, the decoder updates its state vector $\mathbf{s}_i$ and samples a word from distribution $\mathbf{o}_i$ as follows:

$$\begin{aligned} y_i \sim \mathbf{o}_i &= P(y|y_{<i}; X_t \oplus C_t) \\ &= \text{softmax}(\mathbf{W}_o \mathbf{s}_i) \end{aligned} \tag{6}$$

where $\mathbf{W}_o$ is the weight matrix of the decoder. The decoder's state is updated by

$$\mathbf{s}_i = \mathbf{Decoder}(\mathbf{s}_{i-1}, [\mathbf{Att}(\mathbf{X}_t \oplus \mathbf{C}_t; \mathbf{s}_{i-1}); \mathbf{y}_{i-1}]) \tag{7}$$

where $\mathbf{Att}(\mathbf{X}_t \oplus \mathbf{C}_t; \mathbf{s}_{i-1})$ is an attentive read of the encoded input conditioned on state $\mathbf{s}_{i-1}$, typically using attention mechanism [5]; and $\mathbf{y}_{i-1}$ is the vector representation of the previously generated word y_{i-1}.

The formulation of generation-based models mentioned above is auto-regressive in that these models generate a target sequence word by word, each word conditioned on the words that are previously generated. To make the decoding parallelizable, non-autoregressive models based on Transformer have been proposed to generate all the tokens simultaneously [55, 59]. Non-autoregressive modeling factorizes the distribution over a target sequence given a query into a product of conditionally independent per-step distributions, as follows:

$$\mathcal{P}(Y|X_t \oplus C_t) = \prod_{i=1}^{m} P(y_i|X_t \oplus C_t). \tag{8}$$

Though the performance of such non-autoregressive models is still not as good as their autoregressive counterparts, it opens new opportunities for fast training using very large scale datasets [42, 59].

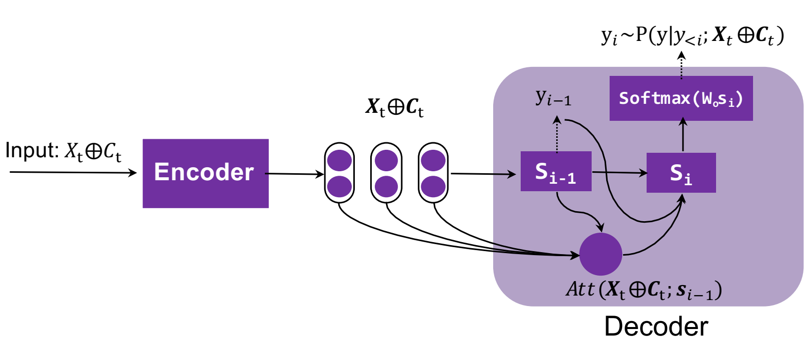

Fig. 4. Typical encoder-decoder framework for generation-based models. The input $X_t \oplus C_t$ is encoded into vectors $\mathbf{X}_t \oplus \mathbf{C}_t$. In the decoder, a word y_i is sampled from $P(y|y_{<i}, X_t \oplus C_t) = softmax(\mathbf{W}_o \mathbf{s}_i)$ and the decoder's state is updated with y_{i-1} and $\mathbf{Att}(\mathbf{X}_t \oplus \mathbf{C}_t; \mathbf{s}_{i-1})$ as input.

Noticeably, the large-scale pre-trained models, such as BERT and GPT-2 [23, 106], can be easily applied in the above encoder-decoder framework. The encoder can be a pre-trained BERT model or a GPT-2 model, the decoder a GPT-2 model. Both the parameters of the encoder and the decoder are initialized using the pre-trained models and then fine-tuned on a dialog corpus [1, 39, 106, 155, 179]. The fine-tuning process is often tailored to the dialog scenario via encoding with dialog state embeddings[155], classifying golden and negatively sampled responses given the same dialog context[39], designing dialog-specific pre-training tasks [11, 84], and so on. These models have shown strong performance in the NeurIPS Conversational Intelligence Challenge 2 (ConvAI 2)[10] and were used in the TREC Conversational Assistance Track (Conversational Information Seeking)[11]. Notably, Zhang et al. [179] released the DialoGPT model that was trained on 147M conversation-like exchanges extracted from on Reddit comment threads, providing a good starting point for future research.

2.3 Hybrid Methods

Retrieval-based methods retrieve an output response from a repository of human-human conversations. Such human-produced conversations are fluent, grammatical, and of high quality. However, the scale of the repository is critical to the success of the methods, which unfortunately is never large enough for open-domain dialog systems. Moreover, retrieval-based methods cannot generate unseen responses. On the other hand, generation-based methods can produce novel responses. But they often generate undesirable responses that are either ungrammatical or irrelevant. Hybrid methods combine the strengths of both and usually adopt a two-stage procedure [153, 165, 192]. In the first stage, some relevant conversations, known as prototype responses in [158], are retrieved from a dataset using input $X \oplus C$ as a query. Then, prototype responses are used to help generate new responses in the second stage.

Based on the Seq2Seq architecture, Song et al. [128] used additional encoders to represent the set of retrieved responses, and applied the attention [5] and copy [43] mechanism in decoding to generate new responses. Pandey et al. [96] first retrieved similar conversations from training data using a TF-IDF model. The retrieved responses were used to create exemplar vectors that were used by the decoder to generate a new response. Wu et al. [158] first retrieved a prototype response from training data and then edited the prototype response according to the differences between the prototype context and current context. The motivation is that the retrieved prototype provides a good start-point for generation because it is grammatical and informative, and the post-editing process further improves the relevance and coherence of the prototype. Zhang et al. [172] proposed an adversarial learning framework to enhance a retrieval-generation ensemble model. Their model consists of a language-model-like generator, a ranker generator, and a ranker discriminator. This model encourages the two generators to generate responses that are scored higher by the discriminative ranker, while the discriminator down-weighs adversarial samples and selects those responses that are favored by the two generators.

3 SEMANTICS

A typical symptom of a dialog system that suffers from the semantics issue is that it often generates bland and generic responses, such as “I don’t know”, “thank you”, “OK” , or simply repeats whatever a user says [33, 120, 129, 141]. We observe similar phenomena in human conversations. When we don’t understand what the other party is talking about but have to respond, we often pick those safe but bland responses.

[10]http://convai.io/
[11]http://www.treccast.ai/

To make an engaging conversation, the dialog system needs to produce contentful, interesting, and interpersonal responses based on its understanding of the dialog content, user's sentiment and emotion, and real-world knowledge that is related to the dialog. In this section, we review some of the most prominent neural approaches that have been proposed recently to address the semantics issue. We first describe the ways of improving the encoder-decoder framework to generate diverse and informative responses by improving the understanding (embedding) of dialog context and users. Then, we describe the methods of grounding dialog in real-world knowledge to make system responses more contentful.

3.1 Improving Diversity and Informativeness in Neural Response Generation

Most state of the art neural response generation models are based on the encoder-decoder framework which consists of four components: (1) an encoder that encodes user input and dialog context, (2) an intermediate representation, (3) an decoder that generates candidate responses, and (4) a ranker that picks the best candidate as the response. In what follows, we review the proposed methods in four categories, each focusing on improving one of the four components.

Encoder. Encoding richer information from query $X \oplus C$, such as longer dialog history [129], persona [62], hidden topics [121], has proved to be helpful for generating more informative responses. Xing et al. [160] extracted topic words, rather than hidden topic vectors, using LDA, and encoded such words in a topic-aware model. The model generates a response by jointly attending to query $X \oplus C$ and the topic words. Topic words are also used to model topic transition in multi-turn conversations [146]. The hybrid methods described in Section 2.3 [96, 128, 158] encode the retrieved prototype responses to help generate more informative responses.

Intermediate Representation. Instead of encoding $X \oplus C$ using a fixed-size vector as in [133], methods have been proposed to use more flexible intermediate representations (e.g., additional latent variables) to enhance the representation capability to address the one-to-many issue in dialog, and to improve the interpretability of the representation in order to better control the response generation. Zhao et al. [185] introduced CVAE for dialogue generation and adopted a Gaussian distribution, rather than a fixed-size vector, as the intermediate representation, thus obtaining more diverse responses via sampling the latent variable. Du et al. [26] introduced a sequence of continuous latent variables to model response diversity, and demonstrated empirically that it is more effective than using a single latent variable. Zhao et al. [184] proposed an unsupervised representation learning method to use discrete latent variables, instead of dense continuous ones, which improves the interpretability of representation. Zhou et al. [187, 188] assumed that there exist some latent responding mechanisms, each of which can generates different responses for a single input post. These responding mechanisms are modeled as latent embeddings, and can be used to encode the input into mechanism-aware context to generate responses with the controlled generation styles and topics. Gao et al. [35] proposed a SpaceFusion model which induces a latent space that fuses the two latent spaces generated by Seq2Seq and auto-encoder, respectively, in such a way that after encoding $X \oplus C$ into a vector in the space, the distance and direction from the predicted response vector given the context roughly match the relevance and diversity, respectively.

Decoder. Assigning additional probability mass to *desirable* words in decoder is a commonly used method to gain some control of what to generate. Mathematically, this can be implemented by adjusting the output word distribution as follows:

$$P_{new}(y_i|y_{<i};X,C) = \mathbf{Normalize}(P(y_i|y_{<i};X,C) + P_{bias}(y_i|y_{<i};X,C)) \tag{9}$$

where $y_{<i} = y_1 y_2 \cdots y_{i-1}$ is the generated prefix; P_{bias} assigns additional probability mass to the desirable words to be generated; and **Normalize**$(\cdot)$ is a normalization function to ensure a probability distribution. Many existing controllable decoding methods essentially fall into this formulation. The most notable example is CopyNet [43], which copies desirable but infrequent words from the input to the output, thus assigning higher probabilities to those words. In [173], P_{bias} is formulated as a Gaussian distribution, which assigns higher probabilities to rare words to control the specificity of a response, where the specificity score of a word is proportional to its IDF (inverse document frequency) score.

Candidate Ranker. To obtain more diverse responses, beam search is commonly used to generate multiple candidates, which (together with retrieved candidates in hybrid dialog systems) are then ranked by another model, which uses information that is not available in decoding (e.g., mutual information between input and response) or is too expensive to use in decoding (e.g., a large pre-trained language model such as BERT [23]) to select the final response. Li et al. [61] proposed to use Maximum Mutual Information (MMI) as the objective to rank candidates to promote the diversity of generated responses. As the standard beam search often produces near-identical results, recent work addresses it by encouraging the diversity among (partial) hypotheses in the beam. For example, Li et al. [63] penalized lower-ranked siblings extended from the same parents, so that the N-best hypotheses in the beam at each time step are more likely to expand from different parents, and thus more diverse. Vijayakumar et al. [140] divided the hypotheses into several groups and applied beam search group-by-group. The model favours the hypotheses that are dissimilar to the ones in the previous groups. Constrained beam search [6] was also proposed to generate desirable responses by constraining a generated response to obey the input structure.

3.2 Knowledge Grounded Dialog Models

Knowledge is crucial for language understanding and generation. To build effective human-machine interactions, it is indispensable to ground the concepts, entities, and relations in text in commonsense knowledge or real-world facts such as those stored in Freebase and Wikipedia. An knowledge-grounded open-domain dialog system should be able to identify the entities and topics mentioned in user input, link them into real-world facts, retrieve related background information, and thereby respond users in a proactive way e.g., by recommending new, related topics to discuss.

Knowledge has been shown useful in both retrieval-based and generation-based dialog systems. A well-known example of the former is Microsoft XiaoIce [192]. XiaoIce relies on a large knowledge graph (KG) to identify the topics and knowledge related to user input for both response generation and topic management. In [168], a Tri-LSTM model is proposed to use commonsense knowledge as external memories to facilitate the model to encode commonsense assertions for response selection. An early example of using knowledge for generating responses is [44], where manually crafted templates are used to generate responses which are filled with relevant knowledge triples. In [36], a knowledge-grounded model is proposed to generate a response by incorporating some retrieved posts that are relevant to the input. However, the quality of these unstructured posts is mixed. Pre-compiled structured knowledge, which is in the form of fact triples, is believed to be of higher quality and has been shown to more helpful in conversation generation [74, 196]. Zhu et al. [196] dealt with a scenario where two speakers are conversing based on each other's private knowledge base in the music domain. The generation model can generate a word in response from the context or the knowledge base. In [74], a knowledge diffusion model is proposed to not only answer factoid questions based on a knowledge base, but also generate an appropriate response containing knowledge base entities that are relevant to the input. Zhou et al. [190] exploited the use of large-scale commonsense knowledge for conversation generation. First, a one-hop subgraph is retrieved from ConceptNet [130] for each word in an input post. Then, the word vectors, along with the graph vectors which extend the meaning of the word via its neighboring entities and relations, are used to encode the input post. During decoding, a graph attention mechanism is applied in which the model first attends to a knowledge graph and then to a triple within each graph, and the decoder chooses a word to generate from either the graph or the common vocabulary. Qin et al. [104] presented a new end-to-end approach that jointly models response generation and on-demand machine reading for generating contentful conversations. The key idea is to provide the model with relevant long-form text on the fly as a source of external knowledge. The model performs QA-style reading comprehension on this text in response to each conversational turn, thereby allowing for more focused integration of external knowledge than prior approaches.

We summarize the aforementioned knowledge-grounded dialog systems in Table 2. Most these studies focus on two problems: (1) *knowledge selection* – selecting appropriate knowledge to be incorporated in the next response given the dialog context and previously-selected knowledge [68, 77, 110, 178], and (2) *knowledge-aware generation* – injecting the required knowledge into a generated response [36, 67, 104, 190]. In addition, zero-shot adaptation to updated, unseen knowledge graphs without conversational data [17] is worth more comprehensive exploration in the future. solving the problem would allow dialog systems to generate proper responses with selected knowledge even though the knowledge has never been used.

Recently, there is a significant burst in constructing document or knowledge grounded dialog corpora [17, 25, 40, 86, 87, 104, 156, 191], which will be described in Section 7 in details.

4 CONSISTENCY

A human-like dialog system needs to embody consistent behaviors, so that it can gain the user's confidence and trust [126, 192]. The consistency issue refers to generating responses that are consistent in persona, style, and context (with respect to topic, logic, causality, etc.). We group existing studies into three lines: (1) persona consistency modeling including implicit and explicit methods, (2) stylistic response generation, and (3) contextual consistency.

4.1 Persona Consistency

Existing dialog models that address persona consistency can be roughly grouped into two catetories: *implicit personalization* and *explicit personalization*. In the former, the persona is implicitly represented by a persona vector. For instance, Kim et al. [57] proposed a ranking-based approach to integrate a personal knowledge base and user interests in dialogue system. Bang et al. [8] extended the user input by exploiting examples retrieved from her personal knowledge base to help identify the candidate responses that fit her persona. Li et al. [62], Zhang et al. [176] used an embedding vector to represent a user (speaker) persona and fed the user embedding into each decoding position of the decoder. Such models need to be trained using conversational data labeled by user identifiers, which is expensive to collect for large quantities. Thus, Wang et al. [144] proposed to train personalized models with only group attributes (e.g., male or female). The group attributes are embedded to vectors and then fed into the decoder for response generation. Zhang et al. [177] proposed a neural conversation model that generates consistent responses by maintaining certain features related to topics and personas throughout the conversation. Unlike other work that requires external supervision such as user identities, which are often unavailable, this approach trains topic and persona feature extractors in a self-supervised way by utilizing the natural structure of dialogue data. Although Ouchi and Tsuboi [93], Zhang et al. [174] showed that user embedding is an effective technique to distinguish roles of speakers and addressees in multi-party conversation, personalization in these models are handled in an implicit way and thus not easy to interpret and control in generating desired responses.

Authors	Knowledge modality	Grounding method	Issues focused	Models
Zhou et al. [190]	Knowledge graph	Retrieval	Knowledge-aware generation	Seq2Seq+Graph Attention
Ghazvininejad et al. [36]	Unstructured text	Retrieval	Knowledge-aware generation	Memory Networks
Zhu et al. [196]	Knowledge base	Retrieval	Knowledge-aware generation	Seq2Seq+ Knowledge Retriever
Liu et al. [74]	Knowledge base	Retrieval	Knowledge-aware generation	HRED+ Knowledge Retriever
Qin et al. [104]	Unstructured text	QA	Knowledge-aware generation	SAN + Generator
Chen and Lee [17]	Knowledge graph	Multi-hop reasoning	Knoweldge-aware generation +Zero-shot adaptation	Seq2Seq+ Multi-hop Reasoning
Dinan et al. [25]	Unstructured text	Retrieval	Knowledge-aware generation	Transformer
Gopalakrishnan et al. [40]	Unstructured text	Retrieval	Knowledge-aware generation	Transformer
Moghe et al. [86]	Unstructured text+ Fact table	Grounding label	Knowledge-aware generation	HRED/GTTP/BiDAF
Moon et al. [87]	Knowledge graph	Grounding label	Knowledge selection	KG path decoder
Wu et al. [156]	Unstructured text+ Knowledge graph	Grounding label	Proactive conversation	BERT/PostKS
Zhou et al. [191]	Unstructured text	Grounding label	Knowledge-aware generation	Seq2Seq
Lian et al. [68]	Unstructured text	Grounding label	Knowledge selection	PostKS
Liu et al. [77]	Unstructured text+ Knowledge graph	QA+Grounding label	Knowledge selection	RL+BiDAF
Ren et al. [110]	Unstructured text	Grounding label	Knowledge selection	BiDAF+GTTP
Zhang et al. [178]	Unstructured text	Grounding label	Knowledge selection	BiDAF+Seq2Seq
Li et al. [67]	Unstructured text	Grounding label	Knowledge-aware generation	Incremental Transformer+ Two-pass Decoder

Table 2. Survey on existing knowledge-grounded studies. Grounding method refers to the means of a grounded knowledge linking to an utterance. Retrieval means that the grounded knowledge is retrieved based on key words in utterances. QA means the knowledge is extracted using machine reading comprehension methods. Grounding label means the knowledge used in the conversation is explicitly annotated by hand. In the last column, PostKS means selecting knowlege by mininizing the KL loss between a prior and a posterior distribution over knowledge [68, 156]; SAN refers to the Stochastic Answer Network for machine reading comprehension model proposed in [76]; GTTP (Get To The Point) refers to the hybrid pointer generator network for abstractive summarization proposed in [115]; HRED refers to the hierarchical neural response generation model [119]; and BiDAF refers to the Bi-Directional Attention Flow network for reading comprehension [117].

In [103], an explicit persona model is proposed to generate personality-coherent responses given a pre-specified user profile. The chatbot's persona is defined by a key-value table (i.e., profile) which consists of name, gender, age, hobbies, and so on. During generation, the model first chooses a key-value from the profile and then decodes a response from the chosen key-value pair forward and backward. This model can be trained on generic dialogue data without user identifier. XiaoIce also uses an explicit persona model [192].

We have discussed two categories of methods for modeling persona consistency: *implicit modeling* [62, 177] which utilizes learned user persona features to capture user-level consistency implicitly, and *explicit modeling* [103, 192] which controls the conversation generation using explicitly-defined user profile. However, most existing methods are insufficient in modeling the user's psychological personality. For instance, we do not yet have a dialog system that can exhibit extrovert or introvert personality. Building such an intelligent dialog system requires breakthroughs in multi-disciplined research on psychology, cognitive, and social science.

4.2 Stylistic Response Generation

Stylistic response generation [92, 143] can be viewed as a form of personalization in conversation. There are two main challenges: how to disentangle content and style in representation, and how to construct training data containing pairs of responses that are of the same content but in different styles. Wang et al. [143] utilized a small-scale stylistic data and proposed a topic embedding model to generate responses in specific styles and topics simultaneously. Oraby et al. [92] demonstrated that it is possible to automate the construction of a parallel corpus where each meaning representation can be realized in different styles with controllable stylistic parameters.

Stylistic conversation generation is closely related to domain adaptation and transfer learning [14, 85, 143, 176]. The idea is to first train a general conversation model on a large corpus in source domain and then to transfer the model to a new speaker or target domain using small amounts of personalized (or stylistic) data in target domain. Casanueva et al. [14] proposed to automatically gather conversations from similar speakers to improve the performance of policy learning of personalized dialogue systems. Zhang et al. [176] proposed a two-phase transfer learning approach, namely *initialization then adaptation*, to generate personalized responses. They also proposed a quasi-Turing test method to evaluate the performance of the generated responses. Yang et al. [167] presented a transfer learning framework similar to Zhang et al. [176], but proposed to use a new adaptation mechanism based on reinforcement learning. Luan et al. [81] proposed a multi-task learning approach where the response generation and utterance representation are treated as two sub-tasks for speaker role adaptation.

4.3 Contextual Consistency

Unlike the studies on persona consistency, the work on modeling contextual consistency is yet to be explored. Early work has focused on better representing dialog contexts [119, 121] using hierarchical models, which can be viewed as *implicit* modeling of contextual consistency. Recently, Welleck et al. [152] and Dziri et al. [27] characterized the contextual consistency as a natural language inference (NLI) problem [22]. In this setting, a response is considered consistent if it can be inferred from the dialog context or the given persona. Welleck et al. [152] constructed a dialog NLI dataset based on Persona-Chat[175]. Zhang et al. [177] proposed to learn topic features from dialog context on-the-fly and utilize controllable response generation techniques to generate topic-consistent responses.

5 INTERACTIVENESS

This issue is mainly about how to optimize the *behaviors and strategies of a dialog system* to maximize long-term user engagement and accomplish long-term, complex goals such as providing emotional comfort, or even psychological counseling [3, 12, 101, 171]. To improve interactiveness, it is important to understand user's emotion and affect, in addition to dialog content, and to optimize the system's behavior and interaction strategy in multi-turn conversations.

5.1 Modeling User Emotion

Emotion perception and expression is vital for building a human-like dialog system. Earlier attempts to building emotional dialog systems are mostly inspired by psychology findings. Those systems are either rule-based or trained on small-scale data, and work well only in a controlled environment. Thanks to the availability of large-scale data and the recent progress on neural conversational AI, many neural response generation models have been proposed to perceive and express emotions in an open-domain dialog setting. Zhou et al. [189] proposed Emotional Chatting Machine (ECM) to generate emotional responses given a pre-specified emotion. ECM consists of three components: (1) an emotion category embedding which is fed into each decoding position, (2) an internal emotion state which assumes that the emotion state decays gradually and finally to zero during decoding, and (3) an external memory which allows the model to choose emotional (e.g., *lovely*) or generic (e.g., *person*) words explicitly at each decoding step. The authors also presented some typical emotion interaction patterns in human-human conversations such as *empathy* and *comfort*, which would inspire more fine-grained design of emotion interaction between human and machine. Asghar et al. [4] developed a method of affective response generation that consists of three components: (1) the affective vectors based on Valence/Arousal/Dominance dimensions [19, 150], which serve as a supplement to word vectors; (2) the affective loss functions which maximize or minimize the affective consistency between a post and a response; and (3) the affective beam search algorithm for seeking affective responses. In [195], a conditional variational autoencoder is proposed to generate more emotional responses conditioned on an input post and some pre-specified emojis. Huber et al. [52] studied how emotion can be grounded in an image to generate more affective conversations. In addition to text, the decoder takes as input the scene, sentiment, and facial coding features extracted from a given image. Recently, an empathetic dialog corpus is developed to facilitate the research on modeling empathetic interactions in conversation [109]. We will present dialog datasets in Section 7.

Controlling the emotion or sentiment has become a popular topic in language generation [37, 50, 105]. In [105], an RNN-based language model is trained on large-scale review data where some neurons are reported to be highly correlated with sentiment expression. Ghosh et al. [37] proposed an affective language model which generates an affective sequence from a leading context. At each decoding position, the model estimates an affective vector of the already generated prefix by keyword spotting using the Linguistic Inquiry and Word Count (LIWC) dictionary [100]. The vector is then used to generate the next word. In [145], to generate the reviews of a particular polarity, the authors proposed a multi-class generative adversarial network which consists of multiple generators for multi-class polarities and a multi-class discriminator.

Despite the research effort reviewed, it is still challenging for a dialog system to express complex emotions in natural language. One difficulty is emotion representation. A simple approach is to project an emotion label to a vector [189], which is implicit, unexplainable, and subtle. A more sophisticated method is to use Valence/Arousal/Dominance representations: the emotion of each word, sentence, and user state can be represented as V-A-D vectors [19, 150], which is intended to capture psychological and linguistic clues beyond the emotion vector. Another issue of most existing work is that the user's emotion transition during a conversation [82] is not explicitly modeled. This is crucial for a dialog system to establish a long-term connection with a user because the user is more willing to engage with the system if the system can always detect negative change of her emotion during the conversation and cheer her up through e.g., shifting to new topics that are more comfortable for both parties.

5.2 Modeling Conversation Behavior and Strategy

As pointed out in [192], an open-domain dialog system needs to have enough social skills to have engaging conversations with users and eventually establish long-term emotional connections with users. These social skills include topic planning and dialog policy which can determine whether to drive the conversation to a new topic when e.g., the conversation has stalled, or whether or not to be actively listening when the user herself is engaged in the conversation. Nothdurft et al. [91] elucidated the challenges of proactiveness in dialogue systems and how they influence the effectiveness of turn-taking behaviour in multimodal and unimodal dialogue systems. Yu et al. [169] proposed several generic conversational strategies (including grounding on entities and OOV words, topic switch, activity initiation, and joke telling) to handle possible system breakdowns in non-task-oriented dialog systems, and designed policies to select these strategies according to dialog context. Zhang et al. [170] addressed the problem of predicting from the very beginning of a conversation whether it will get out of hand. The authors developed a framework for capturing pragmatic devices, such as politeness strategies and rhetorical prompts, used to start a conversation, and analyzed their relation to its future trajectory. Applying this framework in a controlled setting, it is possible to detect early warning signs of antisocial behavior in online discussions.

The above studies inspire researchers to devise new methods of incorporating social skills into an open-domain dialog system. In [65], a retrieval-based method is proposed to first detect the sign of stalemate using rules, and then retrieve responses that contain the entities that are relevant to the input, assuming that a proactive reply should contain the entities that can be triggered from the ones in the input. Yan and Zhao [164] proposed a proactive suggestion method where a look-ahead post for a user is decoded in addition to the system response, conditioned on the context and the previously generated response. The user can use the generated post directly, or type a new one during conversation. Wang et al. [148] argued that asking good questions in conversation is shown to be an important proactive behavior. A typed decoder is proposed to generate meaningful questions by predicting a type distribution over topic words, interrogatives, and ordinary words at each decoding position. The final output distribution is modeled by the type distribution, leading to a strong control over the question to be generated. Rao and Daumé III [108] also argued that question asking is fundamental to communication, and that a good question is the one whose expected answer will be useful. They built a neural network model for ranking clarification questions, evaluated on a dataset of clarification questions (post-question pairs) extracted from StackExchange. Ke et al. [56] conducted a systematic study of generating responses with different sentence functions, such as interrogative, imperative, and declarative sentences. These sentence functions play different roles in conversations. For instance, imperative responses are used to make requests, give directions and instructions, or elicit further interactions while declarative responses make statements or explanations. Tang et al. [135] proposed a new dialog planning task in which the conversation should eventually reach a target (defined by a topical keyword) from any initial topics. In such a task, it is required to plan proactively the topic path to the final target.

There are two important directions for future research. First is the comprehensive investigation of conversation behaviors in human-human dialog. This is still largely ignored, possibly due to the lack of real-world conversations. The dialog data in online forums [170] and psychological counseling [3, 12, 101, 171] are of high value for this research. But the data in a wide variety of scenarios are still in significant shortage. Second is to create a more sophisticated real-world dialog setting for system development and evaluation. Existing work largely targets at modeling atomic strategy in dialog systems, namely, single strategy for emotion interaction [189], topic control [146], question asking [56, 148], and so on. Most of the studies are merely evaluated with the single-turn setting. However, to accomplish more complex social goals such as emotional comfort or counseling, it is necessary to design composite strategies that consider emotion, topic, and proactivity comprehensively in multi-turn conversation. Therefore, there is increasing demand for collecting or constructing more complex dialog data with well-designed task goals, and for developing more sophisticated dialog policy models.

6 OPEN-DOMAIN DIALOG EVALUATION

Evaluating the quality of an open-domain dialog system is challenging because open-domain conversations are inherently open-ended [107]. For example, if a user asks the question "what do you think of Michael Jackson?", there are hundreds of distinct but plausible responses. Evaluation of a dialog system can be performed manually or in an automatic way. In manual evaluation, human judges are hired to assess the generated results in terms of predefined metrics, with well-documented guidelines and exemplars. Evaluation is conducted by either scoring each individual result (point-wise) or comparing two competing results (pair-wise). In some dialog evaluation challenges, manual evaluation is commonly adopted in the final-stage competition [24, 107]. For instance, the second conversational intelligence challenge [24] adopted manual evaluation by paid workers from Amazon Mechanical Turk and unpaid volunteers, and the organizers reported the rating difference between the two user groups: the volunteers' evaluation had relatively fewer good (i.e. long and consistent) dialogues, while paid workers tended to rate the models higher than the volunteers.

Since manual evaluation is expensive, time-consuming, and not always reproducible, automatic evaluation is more frequently used, especially at the early stage of development. For retrieval-based methods, traditional information retrieval evaluation metrics such as precision@k, mean average precision (MAP), and normalized Discounted Cumulative Gain (nDCG) [83] are applicable. For generation-based models, metrics such as perplexity, BLEU [98], and distinct-n [61], are widely used. Perplexity measures how well a probabilistic model fits the data, and is a strong indicator whether the generated text is grammatical. BLEU, adopted from machine translation, measures the lexical overlap between the generated responses and the reference ones. Distinct-n measures the diversity by computing the proportion of unique n-grams in a generated set. However, [73] argued that automatic metrics such as BLEU, ROUGE [70], and METEOR [7] all have low correlation with manual evaluation. But as pointed out in [33], the correlation analysis in [73] is performed at the sentence level while BLEU is designed from the outset to be used as a corpus-level metric. [32] showed that the correlation of string-based metrics (BLEU and deltaBLEU) significantly increases with the units of measurement bigger than a sentence. Nevertheless, in open-domain dialog systems, the same input may have many plausible responses that differ in topics or contents significantly. Therefore, low BLEU (or other metrics) scores do not necessarily indicate low quality as the number of reference responses is always limited in test set. Therefore, there has been significant debate as to whether such automatic metrics are appropriate for evaluating open-domain dialog systems [33].

Recently, trainable metrics for open-domain dialog evaluation have attracted some research efforts. Lowe et al. [78] proposed a machine-learned metric, called ADEM, for open-domain dialog evaluation. They presented a variant of the VHRED model [121] that takes context, user input, gold and system responses as input, and produces a qualitative score between 1 and 5. The authors claimed that the learned metric correlates better with human evaluation than BLEU and ROUGE.

[136] proposed an evaluation model, called RUBER, which does not rely on human judged scores. RUBER consists of a referenced component to measure the overlap between a system response and a reference response, and an unreferenced component to measure the correlation between the system response and the input utterance. However, as pointed out in [114], ADEM can be easily fooled with a variation as simple as reversing the word order in the text. Their experiments on several such adversarial scenarios draw out counter-intuitive scores on the dialogue responses. In fact, any trainable metrics lead to potential problems such as overfitting and "gaming of the metric" [12] [2], which might explain why none of the previously proposed machine-learned evaluation metrics [2, 20, 38, 58, 72, 94, 131, etc.] is used in official machine translation benchmarks. Readers refer to [33] for a detailed discussion.

These research attempts indicate that what makes a good conversation is a challenging question to answer. See et al. [116] discussed four attributes that are associated with the control of open-domain dialog generation: repetition, specificity, response-relatedness, and question-asking. They argued that existing work has ignored the importance of the conversational flow, because existing models repeat or contradict previous statements, fail to balance specificity with genericness, and are unable to balance asking questions with other dialogue acts. Experiments on Persona-Chat [175] show that higher engagingness scores in human judgement can be obtained by optimizing the control of the four attributes in multi-turn conversations. Therefore, considering these attributes in automatic evaluation, implicitly or explicitly, is expected to lead to new evaluation metrics that correlate well with human evaluation.

Recently, there are research attempts to combine human evaluation and automatic evaluations for natural language generation systems. Hashimoto et al. [45] argued that human evaluation captures quality but not diversity while statistical evaluation (i.e., perplexity) captures diversity but not quality. They proposed a unified framework which evaluates both in terms of the optimal error rate of predicting whether a sentence is human- or machine-generated. As mentioned above, automatic metrics such as sentence-level BLEU correlates poorly with human judgement, thereby easily leading to systematic bias against model improvements. On the other hand, the average of human judgements is unbiased but is very expensive to collect. Therefore, Chaganty et al. [15] combined automatic metrics with human evaluation to obtain an unbiased estimator with lower cost than using solely human evaluation.

All of the above research suggests that automatic evaluation of dialog systems is by no means a solved problem. We argue that, for open-domain dialog evaluation, the major difficulty derives from in the one-to-many essence: in any given dataset, the number of observable responses for the same input post is limited, yet there are many appropriate responses not presented in the dataset. Therefore, automatic metrics that are trained on a dataset will be inherently questionable because the topic coverage and the number of observable outputs are largely limited by the dataset. Thus, uncovering those underlying outputs for an input post is an interesting area for future research.

7 OPEN-DOMAIN DIALOG CORPORA

Recently, the availability of dialog corpora has largely advanced the development of neural models for open-domain conversation generation. An incomplete survey on these dialog datasets is

[12] In discussing the potential pitfalls of machine-learned evaluation metrics, Albrecht and Hwa [2] argued for example that it would be "prudent to defend against the potential of a system gaming a subset of the features." In the case of deep learning, this gaming would be reminiscent of making non-random perturbations to an input to drastically change the network's predictions, as it was done, e.g., with images in [134] to show how easily deep learning models can be fooled. Readers refer to Chapter 5 in Gao et al. [33] for a detailed discussion.

presented in Table 3[13]. These corpora differ in topic, source (where or how the data is collected), language, data scale, and the design features.

Short Text Conversation (STC) [123]: This corpus is collected from a Chinese social media, Weibo. There are 219,905 posts and 4,308,211 responses in the training data. It can be used for studying the one-to-many problem in dialog modeling since each post has multiple responses. On top of this corpus, Zhou et al. [189] proposed an emotional STC dataset (ESTC) in which each utterance is tagged in terms of six emotion classes by an emotion classifier with an accuracy of 62.3%. ESTC is frequently used in building empathetic dialog systems [189].

Twitter Triple Corpus [129]: This corpus contains 29M context-message-response triples from the Twitter FireHose, covering the 3-month period from June 2012 through August 2012. Additionally, the validation and test sets have 4,232 triples which are scored no less than 4 in 5-point scale by human annotators. However, this corpus is not publicly available.

PersonalDialog [186]: This corpus is constructed toward building personalized conversation models. The data is collected from a Chinese social media, Weibo. Each dialogue is composed of a post and its following replies from different users. The personal profile of each user is collected, which includes five personality traits: Gender, Age, Location, Interest Tags, and Self Description. This dataset contains 20.83M conversations and 8.47M user profiles. The total number of utterances are 56.25M and each utterance contains 9.35 tokens. A considerable amount of dialogues (3.43M sessions) in this dataset have multiple turns (more than 4 utterances). This corpus is the first dialogue corpus that contains real social conversations and diversified personality traits for each user.

DailyDialog [66]: This corpus contains multi-turn dialogs on daily life topics. The raw data were crawled from several websites which serve for English learner to practice English. The dataset contains 13,118 dialogs, with an average of 7.9 turns per dialog and 14.6 words per turn. The appealing feature of this corpus is that it provides manual annotation on intent (*Inform, Questions, Directives, and Commissive*) and emotion (*Anger, Disgust, Fear, Happiness, Sadness, and Surprise*), which may support the research on emotion interaction and dialog act modeling.

Ubuntu Dialog Corpus [79]: This corpus contains two-party conversations that solve technical issues with Ubuntu. The data were extracted from online conversation logs in Ubuntu-related chat rooms on the Freenode Internet Relay Chat (IRC) network. In each log, a user may ask a technical question to be solved and other users can respond to the question. The log session will terminate until the problem is solved. A two-party conversation will be extracted from the chat log[14]. The corpus contains 930,000 human-human dialogs and 7,100,000 utterances, with an average of 7.71 turns per dialog and 10.34 words per utterance. Strictly speaking, this dataset is task-specific instead of open-domain conversation. This corpus is commonly used to evaluate retrieval-related models.

Persona-Chat [175]: This crowdsourced corpus is designed for personalized dialog modeling. In each conversation, each worker is given a persona which is defined by up to 5 sentences describing personal hobby or state (e.g., *I like swimming*, or *I need to lose weight*). Two workers are instructed to know each other through interaction. During the conversation, each worker should follow her own persona and try to know the partner's information. The dataset consists of 10,981 dialogs with 164,356 utterances.

CMU Document-grounded conversation (CMU DOG) [191]: This corpus, designed for document or knowledge grounded dialog modeling, contains crowd-sourced conversations that are

[13]Readers may refer to an old survey published in 2015, which covers datasets for both open-domain and task-oriented dialog models [118]. We only list the corpora that are frequently used or recently proposed in the literature, most of which are not covered by [118].

[14]Each chat log is a multi-party conversation, but only two-party sub-conversations which involve the same two users are retained.

talking about 30 movies. The information about each movie is given through a correspondent Wikipedia article. There are two modes for data collection: only one worker has the movie document and both workers have the movie document during conversation. The dataset consists of 4,112 conversations with an average of 31.6 utterances per dialog and 10.8 words per utterance.

[86]: The corpus can be viewed as an expanded version of CMU DOG. The conversations discuss about 921 movies, and the knowledge about each movie is composed of a fact table[15], the plot description, and reviews and comments on the movie. The corpus contains 9,071 conversations and 90,810 utterances with an average of 10 utterances per dialog and 15.3 words per utterance. The corpus is useful for studying the use of heterogeneous knowledge in conversation generation.

Wizard of Wikipedia [25]: This corpus contains conversations that are grounded with knowledge retrieved from Wikipedia. The dataset covers 1,365 topics, each linked to a Wikipedia article. These topics include commuting, Gouda cheese, music festivals, podcasts, bowling, and Arnold Schwarzenegger. Each conversation is made between a knowledge expert and a curious learner, and the expert has full access to the Wikipedia article of a topic but the learner does not. The corpus consists of 22,311 dialogues and 201,999 utterances, with an average of 9 utterances per dialog. Each utterance is grounded to a selected knowledge sentence or indicated by that no knowledge is used.

Grounded Response Generation at DSTC7 [104]: The dataset, which is first released for the "sentence generation" task at the 7th Dialog System Technology Challenges (DSTC7) [31], is developed for grounded conversation modeling. It consists of conversation threads extracted from Reddit data. Each conversation contains exactly one URL to a web page (grounding) that defines the topic of the conversation. The dataset contains 2.8M conversation instances respectively divided into train, validation, and test based on date ranges: years 2011-2016 for train, Jan-Mar 2017 for validation, and the rest of 2017 for test, which consists of 2,208 conversational turns, each with 6 human responses. To access the human performance using the test set, one of the 6 human responses is set aside, and the remaining 5 responses serve as ground truths for evaluating different systems.

Topical-Chat [40]: This corpus is designed towards building dialog systems that can converse with humans on various topics. It covers 300 popular topic entities spanning 8 domains including fashion, politics, books, and sports. For each entity, the authors fetched the Wikipedia lead section, and crowdsourced 8-10 fun facts. Furthermore, they fetched Washington Post articles in 2018 that each referenced 3 or more of the 300 entities. The authors then created a set of reading sets, each containing the wiki-information, several fun facts, and a Washington Post article. Workers were partnered up to converse, with symmetric or asymmetric settings where symmetric means two workers have the same reading set, and asymmetric with different sets. The dataset contains 11,319 conversations and 248,014 utterances with an average of 22 turns per dialog and 19.8 words per turn.

OpenDialKG [87]: In this corpus, each dialog is paired with its corresponding "knowledge graph (KG) paths" that weave together the KG entities and relations. It was collected with a Wizard-of-Oz setting by connecting two crowd-workers to engage in a chat session. The first worker is given a seed entity and asked to initiate a conversation about that entity. The second worker is provided with a list of facts relevant to that entity, and asked to choose the most natural and relevant facts and use them to frame a free-form conversational response. After the second worker sends her response, new multi-hop facts from KG are surfaced to include paths initiating from new entities introduced in the latest message. The circle continues for several rounds, which simulates a random walk over the knowledge graph. The dataset covers four domains (movies, books, sports and music),

[15] which contains box office collection, similar movies (for recommendation), awards, and tag-lines.

with a KG of total 1,190,658 fact triples. It contains 15,673 dialogs and 91,209 turns with an average of 5.8 turns per dialog. This corpus is useful in studying conversational reasoning, while it is not yet publicly available.

DuConv [156]: This corpus covers topics on movies and film stars whose related knowledge was crawled from the Web. Then, two linked entities were randomly sampled to construct a conversation goal like "[start]$\rightarrow$ $entity_a \rightarrow entity_b$" where $entity_b$ is the final target of the conversation. Two annotators were asked to conduct knowledge-driven conversations with a leader-follower mode. The leader needs to change the conversation topics following the conversation goal and meanwhile keeps the conversation as engaging as possible. The dataset contains 29,858 dialogs and 270,399 utterances with an average of 9.1 turns per dialog and 10.6 words per turn. This corpus is useful in constructing knowledge-driven proactive dialogue systems.

DyKgChat [17]: This corpus is collected for knowledge-grounded conversation modeling. The conversations are from the scripts of a Chinese palace drama (Hou Gong Zhen Huan Zhuang, with 76 episodes and hundreds of characters), and an English sitcom "Friends" (with 236 episodes and six main characters). The paired knowledge graphs are manually constructed. The corpus contains 1,247/3,092 dialogs, with 13.76/18.68 turns per dialog and 27.0/16.5 words per turn for the Chinese and English TV series, respectively. The most interesting feature of this corpus is that it contains evolving knowlege graphs.

EmpatheticDialogues [109]: This corpus is constructed toward building empathetic open-domain conversation models. The data is collected by crowd workers with a speaker-listener mode. The speaker starts the conversation from a pre-set emotion state (e.g., *afraid*) and a personal situation description (e.g., *Speaker felt afraid when she has been hearing noises around the house at night*), and the listener becomes aware of the underlying situation through what the Speaker says and responds. The corpus contains 24,850 conversations, and the average number of utterances per conversation and words per turn is 4.31/15.2 respectively. The corpus is useful in modeling emotion interactions in multi-turn conversation.

Target-Guided Conversation [135]: This corpus is constructed towards building target-guided open-domain conversation models. It's derived from Persona-Chat [175] without the persona information. The keywords of each utterance, which indicate the targets in this task, are automatically extracted by a rule-based keyword extractor. The corpus contains 8,939/500/500 dialogs, 101,935/5,602/5,317 utterances and 2,678/2,080/1,571 keywords in the training/validation/test set, respectively. The average number of keywords in each utterance is about 2.0. This corpus is expected to model the turn-level keyword transition and the discourse-level target-guided dialogue strategy.

PERSUASION-FOR-GOOD [147]: This corpus contains persuasion conversations for charity donation where each speaker's psychological profile attributes and sociodemographic backgrounds such as age and income were also collected. The data is collected with a persuader-persuadee mode in four steps. First, workers were asked to complete a pre-task survey to assess their psychological profile variables. Second, two workers were randomly assigned the roles of persuader and persuadee where the persuader needed to persuade the persuadee to donate part of his/her task earning to the charity, and the persuader could also choose to donate. Third, both the persuader and the persuadee were asked to input the intended donation amount privately though a text box when the conversation was ended. Last, workers were asked to complete a post-survey to assess their sociodemographic backgrounds. The corpus contains 1,017 dialogs, with an averge of 10.43 turns per dialog and 19.36 words per utterances. It also provides manual annotation in terms of persuasion strategy and dialog act for each sentence. This dataset is interesting for studying personalized dialog and complex strategy modeling.

8 DISCUSSIONS AND FUTURE TRENDS

In this paper, we review the recent progress in developing open-domain dialog systems. We focus the discussion on neural approaches that have been proposed to deal with three key challenges: semantics, consistency, and interactiveness. We review open-domain dialog evaluation metrics for both manual and automatic evaluation, and share our thoughts on how to develop better automatic evaluation metrics. We survey frequently-used and recently-proposed corpora for the development of evaluation of open-domain dialog systems.

Differing from early generations of dialog assistants which are designed for simple tasks that require only short, domain-specific conversations, such as making reservation or asking for information, open-domain dialog systems are design to be AI companions that are able to have long, free-form social chats with human users. [107, 192]. Despite the recent progress as reviewed in this paper, achieving sustained, coherent, and engaging open-domain conversations remains very challenging. We conclude this paper by discussing some future research trends

Topic and Knowledge Grounding. To deliver contentful conversations, it is important to ground conversations in real-world topics and entities (e.g., in knowledge bases). This is part of the semantics challenge we have discussed in Section 3. Since natural language understanding in open-domain dialog systems is extremely challenging, knowledge grounding provides to some degree the ability of understanding language in dialog context, as shown in several preliminary studies [74, 190, 196]. Even though an open-domain dialog system has no access to annotated dialog acts (which are available only for task-oriented dialog) to learn to explicitly detect an user's intents (labeled by dialog acts), the system can still play a proactive role of leading the conversation by for example suggesting new topics, if the key concepts and entities are correctly recognized and linked to a knowledge base [30, 102, 148, 192]. Several recently proposed corpora, as described in Section 7, provide new test beds for this research.

Empathetic Computing. Sentiment and emotion form a key factor for making effective social interactions, and is crucial for building an empathetic social bot. Existing studies [4, 109, 189, 192, 195] in this direction are still in the infant stage, as they only deal with superficial expression of emotion. A future empathetic machine should be able to perceive a user's emotion state and change, deliver emotionally influential conversations, and evaluate the emotional impact of its action, much of which should be tightly aligned with psychological studies. These become more important in more complicated scenarios such as psychological treatment, mental health, and emotional comforting. Moreover, it is insufficient for an empathetic machine to use only text information. The signals from other modalities such as facial expression and speech prosody should also be leveraged [21, 69, 182]. To foster the research, Saha et al. [113] developed a conversational dataset consisting of multi-modal dialog sessions in a fashion domain where each turn contains a textual utterance, one or more images, or a mix of text and images.

Personality of a Social Bot. A coherent personality is important for a social bot to gain human trust, thereby improving the consistency and interactiveness of human-machine conversations. Personality (e.g., Big five traits) has been well-defined in psychology [41, 90]. However, existing studies [62, 103, 175, 192] are yet to be significantly extended by incorporating the results of multidiscipline research covering psychology, cognitive science, computer science, etc. The central problem is how to ensure personality-coherent behaviors in conversations and evaluate such behaviors from the perspectives of multidisciplines, particularly via psychological studies.

Controllability of dialog generation. Most existing open-domain dialog systems are based on neural response generation models. Due to the essence of probabilistic sampling used in language generation, controllability is a challenging issue as repetitive, bland, illogical or even unethical responses are frequently observed. Controllability is closely related to the interpretability and robustness of neural network models. Achieving controllability requires new breakthroughs in modeling, such as the hybrid approaches that combine the strengths of both neural and symbolic methods.

9 ACKNOWLEDGEMENT

This work was supported by the National Science Foundation of China (Grant No. 61936010/61876096), and the National Key R&D Program of China (Grant No. 2018YFC0830200). We would like to thank THUNUS NExT Joint-Lab for the support.

We would like to thank Pei Ke, Qi Zhu, Chujie Zheng, Yaoqin Zhang, Hao Zhou, Chris Brockett, Bill Dolan, and Michel Galley for their discussions and contributions to this paper. We truly thank anonymous reviewers for their valuable reviews and comments.

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[2] Joshua Albrecht and Rebecca Hwa. 2007. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of ACL 2007, June 23-30, 2007, Prague, Czech Republic.* 880–887.

[3] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics* 4 (2016), 463–476.

[4] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective Neural Response Generation. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, Vol. 10772. 154–166.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015, San Diego, CA, USA, May 7-9, 2015.*

[6] Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. 2019. Constrained Decoding for Neural NLG from Compositional Representations in Task-Oriented Dialogue. *arXiv preprint arXiv:1906.07220* (2019).

[7] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.* 65–72.

[8] Jeesoo Bang, Hyungjong Noh, Yonghee Kim, and Gary Geunbae Lee. 2015. Example-based chat-oriented dialogue system with personalized long-term memory. In *2015 International Conference on Big Data and Smart Computing, BIGCOMP 2015, Jeju, South Korea, February 9-11, 2015.* 238–243.

[9] Antoine Bordes, Y.-Lan Boureau, and Jason Weston. 2017. Learning End-to-End Goal-Oriented Dialog. In *ICLR 2017, Toulon, France, April 24-26, 2017.*

[10] Basma El Amel Boussaha, Nicolas Hernandez, Christine Jacquin, and Emmanuel Morin. 2019. Deep Retrieval-Based Dialogue Systems: A Short Review. *arXiv preprint arXiv:1907.12878* (2019).

[11] Paweł Budzianowski and Ivan Vulić. 2019. Hello, It's GPT-2–How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. *CoRR* abs/1907.05774 (2019).

[12] Jie Cao, Michael Tanana, Zac E Imel, Eric Poitras, David C Atkins, and Vivek Srikumar. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. *arXiv preprint arXiv:1907.00326* (2019).

[13] Rollo Carpenter. 2011. Cleverbot.

[14] Iñigo Casanueva, Thomas Hain, Heidi Christensen, Ricard Marxer, and Phil D. Green. 2015. Knowledge transfer between speakers for personalised dialogue management. In *Proceedings of SIGDIAL 2015, September 2-4, 2015, Prague, Czech Republic.* 12–21.

[15] Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018.* 643–653.

[16] Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, et al. 2018. Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data. *Alexa prize proceedings* (2018).

[17] Yi-Lin Tuan Yun-Nung Chen and Hung-yi Lee. 2019. DyKgChat: Benchmarking Dialogue Generation Grounding on Dynamic Knowledge Graphs. In *Proceedings of EMNLP 2019, Hong Kong, China, November 3-7, 2019*.

[18] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial Intelligence* 2, 1 (1971), 1–25.

[19] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. In *NAACL 2019, Minneapolis, USA, June 2-7, 2019*. 3734–3743.

[20] Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2001, July 9-11, 2001, Toulouse, France*. 148–155.

[21] Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User Attention-guided Multimodal Dialog Systems. (2019).

[22] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*. Springer, 177–190.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL* (2019), 4171–4186.

[24] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The Second Conversational Intelligence Challenge (ConvAI2). *CoRR* abs/1902.00098 (2019).

[25] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational agents. *CoRR* abs/1811.01241 (2018).

[26] Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. 2018. Variational Autoregressive Decoder for Neural Response Generation. In *Proceedings of EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018*. 3154–3163.

[27] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Evaluating Coherence in Dialogue Systems using Entailment. In *NAACL 2019, Minneapolis, USA, June 2-7, 2019*. 3806–3812.

[28] Yixing Fan, Liang Pang, Jianpeng Hou, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2017. MatchZoo: A Toolkit for Deep Text Matching. *CoRR* abs/1707.07270 (2017).

[29] Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A Smith. 2017. Sounding board–university of washington's alexa prize submission. *Alexa prize proceedings* (2017).

[30] Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. 2018. Sounding Board: A User-Centric and Content-Driven Social Chatbot. In *Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, USA, June 2-4, 2018, Demonstrations*. 96–100.

[31] Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded Response Generation Task at DSTC7. In *AAAI Dialog System Technology Challenges Workshop*.

[32] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. In *Proceedings of ACL-IJCNLP 2015, July 26-31, 2015, Beijing, China*. 445–450.

[33] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural approaches to conversational AI. *Foundations and Trends® in Information Retrieval* 13, 2-3 (2019), 127–298.

[34] Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng. 2014. Modeling interestingness with deep neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2–13.

[35] Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly Optimizing Diversity and Relevance in Neural Response Generation. *arXiv preprint arXiv:1902.11205* (2019).

[36] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A Knowledge-Grounded Neural Conversation Model. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018*. 5110–5117.

[37] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In *Proceedings of ACL 2017, Vancouver, Canada, July 30-August 4, 2017*. 634–642.

[38] Jesús Giménez and Lluís Màrquez. 2008. A Smorgasbord of Features for Automatic MT Evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*. 195–198.

[39] Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-Scale Transfer Learning for Natural Language Generation. In *ACL*. 6053–6058. https://www.aclweb.org/anthology/P19-1608

[40] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain

Conversations. *Proc. Interspeech 2019* (2019), 1891–1895.

[41] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[42] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2017. Non-Autoregressive Neural Machine Translation. *CoRR* abs/1711.02281 (2017). arXiv:1711.02281 http://arxiv.org/abs/1711.02281

[43] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *Proceedings of ACL 2016, Berlin, Germany, August 7-12, 2016.*

[44] Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of SIGDIAL 2015, September 2-4, 2015, Prague, Czech Republic.* 129–133.

[45] Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying Human and Statistical Evaluation for Natural Language Generation. In *NAACL 2019, Minneapolis, USA, June 2-7, 2019.* 1689–1701.

[46] Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Ivan Vulić, et al. 2019. ConveRT: Efficient and Accurate Conversational Representations from Transformers. *arXiv preprint arXiv:1911.03688* (2019).

[47] Matthew Henderson, Blaise Thomson, and Steve J. Young. 2013. Deep Neural Network Approach for the Dialog State Tracking Challenge. In *Proceedings of SIGDIAL 2013, August 22-24, 2013, SUPELEC, Metz, France.* 467–471.

[48] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *COLING 2014, August 23-29, 2014, Dublin, Ireland.* 928–939.

[49] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In *NIPS 2014, Montreal, Quebec, Canada, December 8-13, 2014.* 2042–2050.

[50] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward Controlled Generation of Text. In *Proceedings of ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Vol. 70. 1587–1596.

[51] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013.* 2333–2338.

[52] Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional Dialogue Generation using Image-Grounded Language Models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018.* 277.

[53] Sina Jafarpour, Christopher JC Burges, and Alan Ritter. 2010. Filter, rank, and transfer the knowledge: Learning to chat. *Advances in Ranking* 10 (2010), 2329–9290.

[54] Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988* (2014).

[55] Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast Decoding in Sequence Models Using Discrete Latent Variables. In *Proceedings of ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Vol. 80. 2395–2404.

[56] Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. 2018. Generating Informative Responses with Controlled Sentence Function. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018.* 1499–1508.

[57] Yonghee Kim, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2014. Acquisition and Use of Long-Term Memory for Personalized Dialog Systems. In *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction - Second International Workshop, MA3HMI 2014, Held in Conjunction with INTERSPEECH 2014, Singapore, Singapore, September 14, 2014*, Vol. 8757. 78–87.

[58] Alex Kulesza and Stuart M. Shieber. 2004. A Learning Approach to Improving Sentence-Level MT Evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation.* Baltimore, MD.

[59] Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement. In *Proceedings of EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018.* 1173–1182.

[60] Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *Ai Magazine* 32, 2 (2011), 42–56.

[61] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL HLT 2016, San Diego California, USA, June 12-17, 2016.* 110–119.

[62] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of ACL 2016, Berlin, Germany, August 7-12, 2016.*

[63] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A Simple, Fast Diverse Decoding Algorithm for Neural Generation. *CoRR* abs/1611.08562 (2016).

[64] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation. In *Proceedings of EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017.* 2157–2169.

[65] Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. 2016. StalemateBreaker: A Proactive Content-Introducing Approach to Automatic Human-Computer Conversation. In *Proceedings of IJCAI 2016, New York, NY, USA, 9-15 July 2016*. 2845–2851.

[66] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 986–995.

[67] Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In *Proceedings of ACL 2019, Florence, Italy, July 28 - Aug 2, 2019*. 12–21.

[68] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911* (2019).

[69] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware Multimodal Dialogue Systems. In *2018 ACM Multimedia Conference on Multimedia Conference*. 801–809.

[70] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004).

[71] Zachary C. Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. 2018. BBQ-Networks: Efficient Exploration in Deep Reinforcement Learning for Task-Oriented Dialogue Systems. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018*. 5237–5244.

[72] Lucian Vlad Lita, Monica Rogati, and Alon Lavie. 2005. BLANC: Learning Evaluation Metrics for MT. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (Vancouver, British Columbia, Canada) *(HLT '05)*. 740–747.

[73] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 2122–2132.

[74] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge Diffusion for Neural Dialogue Generation. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018*. 1489–1498.

[75] Tie-Yan Liu. 2010. Learning to rank for information retrieval. In *Proceeding of SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*. 904.

[76] Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic Answer Networks for Machine Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1694–1704.

[77] Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge Aware Conversation Generation with Reasoning on Augmented Graph. *arXiv preprint arXiv:1903.10245* (2019).

[78] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of ACL 2017, Vancouver, Canada, July 30-August 4, 2017*. 1116–1126.

[79] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 285–294.

[80] Zhengdong Lu and Hang Li. 2013. A Deep Architecture for Matching Short Texts. In *NIPS 2013, December 5-8, 2013, Lake Tahoe, Nevada, United States*. 1367–1375.

[81] Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-Task Learning for Speaker-Role Adaptation in Neural Conversation Models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*. 605–614.

[82] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of AAAI 2019, Honolulu, Hawaii, USA, January 27-February 1, 2019*. 6818–6825.

[83] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval.* Cambridge University Press.

[84] Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining Methods for Dialog Context Representation Learning. In *ACL*. 3836–3845. https://www.aclweb.org/anthology/P19-1373

[85] Kaixiang Mo, Yu Zhang, Shuangyin Li, Jiajun Li, and Qiang Yang. 2018. Personalizing a Dialogue System With Transfer Reinforcement Learning. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018*. 5317–5324.

[86] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2322–2332.

[87] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *Proceedings of ACL 2019, Florence, Italy, July 28 - Aug 2, 2019*. 845–854.

[88] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. In *IJCNLP*. 462–472.

[89] Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *Proceedings of ACL 2017, Vancouver, Canada, July 30-August 4, 2017*. 1777–1788.

[90] Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology* 66, 6 (1963), 574.

[91] Florian Nothdurft, Stefan Ultes, and Wolfgang Minker. 2015. Finding appropriate interaction strategies for proactive dialogue systems-an open quest. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, August 6-8, 2014, Tartu, Estonia*. 73–80.

[92] Shereen Oraby, Lena Reed, Shubhangi Tandon, Sharath T. S., Stephanie M. Lukin, and Marilyn A. Walker. 2018. Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators. In *Proceedings of SIGDIAL 2018, July 12-14, 2018, Melbourne, Australia*. 180–190.

[93] Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and Response Selection for Multi-Party Conversation. In *Proceedings of EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 2133–2143.

[94] Sebastian Pado, Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Measuring Machine Translation Quality as Semantic Equivalence: A Metric Based on Entailment Features. *Machine Translation* (2009), 181–193.

[95] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab K. Ward. 2016. Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *IEEE/ACM Trans. Audio, Speech & Language Processing* 24, 4 (2016), 694–707.

[96] Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. 2018. Exemplar Encoder-Decoder for Neural Conversation Generation. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018*. 1329–1338.

[97] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text Matching as Image Recognition. In *Proceedings of AAAI 2016, February 12-17, 2016, Phoenix, Arizona, USA*. 2793–2799.

[98] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002, July 6-12, 2002, Philadelphia, PA, USA*. 311–318.

[99] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Çelikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning. In *Proceedings of EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 2231–2240.

[100] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[101] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1426–1435.

[102] Jan Pichl, Petr Marek, Jakub Konrád, Martin Matulík, Hoang Long Nguyen, and Jan Sedivý. 2018. Alquist: The Alexa Prize Socialbot. *CoRR* abs/1804.06705 (2018).

[103] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation. In *Proceedings of IJCAI 2018, Stockholm, Sweden, July 13-19, 2018*. 4279–4285.

[104] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading. *ACL* (2019).

[105] Alec Radford, Rafal Józefowicz, and Ilya Sutskever. 2017. Learning to Generate Reviews and Discovering Sentiment. *CoRR* abs/1704.01444 (2017).

[106] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Blog* (2018).

[107] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: The Science Behind the Alexa Prize. *CoRR* abs/1801.03604 (2018).

[108] Sudha Rao and Hal Daumé III. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2737–2746.

[109] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of ACL 2019, Florence, Italy, July 28 - Aug 2, 2019*. 5370–5381.

[110] Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation. *arXiv preprint arXiv:1908.09528* (2019).

[111] Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of EMNLP 2011, John McIntyre Conference Centre, Edinburgh, UK, 27-31 July 2011*. 583–593.

[112] Lina Maria Rojas-Barahona, Milica Gasic, Nikola Mrksic, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017*. 438–449.

[113] Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2017. Multimodal Dialogs (MMD): A large-scale dataset for studying multimodal domain-aware conversations. *arXiv preprint arXiv:1704.00200* (2017).

[114] Ananya Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Response Generation by Context-aware Prototype Editing. In *Proceedings of AAAI 2019, Honolulu, Hawaii, USA, Jan 27-Feb 1, 2019*.

[115] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1073–1083.

[116] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1702–1723.

[117] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* (2016).

[118] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742* (2015).

[119] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical Neural Network Generative Models for Movie Dialogues. *arXiv preprint arXiv:1507.04808* (2015).

[120] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of AAAI 2016, February 12-17, 2016, Phoenix, Arizona, USA*. 3776–3784.

[121] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*. 3295–3301.

[122] Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *Proceedings of SIGIR 2015, Santiago, Chile, August 9-13, 2015*. 373–382.

[123] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of ACL 2015, July 26-31, 2015, Beijing, China*. 1577–1586.

[124] Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving Variational Encoder-Decoders in Dialogue Generation. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018*. 5456–5463.

[125] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*. 101–110.

[126] Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of IT & EE* 19, 1 (2018), 10–26.

[127] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *NIPS 2015, December 7-12, 2015, Montreal, Quebec, Canada*. 3483–3491.

[128] Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. 2018. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. In *Proceedings of IJCAI 2018, Stockholm, Sweden, July 13-19, 2018*. 4382–4388.

[129] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *NAACL HLT 2015, Denver, Colorado, USA, May 31 - June 5, 2015*. 196–205.

[130] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of AAAI 2017, San Francisco, California, USA., February 4-9, 2017*. 4444–4451.

[131] Miloš Stanojević and Khalil Sima'an. 2014. Fitting Sentence Level Translation Evaluation with Many Dense Features. In *Proceedings of EMNLP 2014, Doha, Qatar, October 25-29, 2014*. 202–206.

[132] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems. In *Proceedings of ACL 2016, Berlin, Germany, August 7-12, 2016*.

[133] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS 2014, Montreal, Quebec, Canada, December 8-13, 2014*. 3104–3112.

[134] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR* abs/1312.6199 (2013).

[135] Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. 2019. Target-Guided Open-Domain Conversation. In *Proceedings of ACL 2019, Florence, Italy, July 28 - Aug 2, 2019*. 5624–5634.

[136] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018*. 722–729.

[137] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Multi-Representation Fusion Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 267–275.

[138] Alan M Turing. 1950. Computing machinery and intelligence. *Mind* 59, 236 (1950), 433–460.

[139] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[140] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse Beam Search for Improved Description of Complex Scenes. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018*. 7371–7379.

[141] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *CoRR* abs/1506.05869 (2015).

[142] Richard S Wallace. 2009. The anatomy of ALICE. In *Parsing the Turing Test*. Springer, 181–210.

[143] Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering Output Style and Topic in Neural Response Generation. In *Proceedings of EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 2140–2150.

[144] Jianan Wang, Xin Wang, Fang Li, Zhen Xu, Zhuoran Wang, and Baoxun Wang. 2017. Group Linguistic Bias Aware Neural Response Generation. In *Proceedings of the 9th SIGHAN Workshop on Chinese Language Processing, SIGHAN@IJCNLP 2017, Taipei, Taiwan, December 1, 2017*. 1–10.

[145] Ke Wang and Xiaojun Wan. 2018. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In *Proceedings of IJCAI 2018, Stockholm, Sweden, July 13-19, 2018*. 4446–4452.

[146] Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. 2018. Chat More: Deepening and Widening the Chatting Topic via A Deep Model. In *SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. 255–264.

[147] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of ACL 2019, Florence, Italy, July 28 - Aug 2, 2019*. 5635–5649.

[148] Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018*. 2193–2203.

[149] Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral Multi-Perspective Matching for Natural Language Sentences. In *Proceedings of IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 4144–4150.

[150] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.

[151] Joseph Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[152] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671* (2018).

[153] Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and Refine: Improved Sequence Generation Models For Dialogue. *EMNLP 2018* (2018), 87.

[154] Genta Indra Winata, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. 2017. Nora the empathetic psychologist. In *Proc. Interspeech*. 3437–3438.

[155] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2018. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. In *NIPS2018 CAI Workshop*.

[156] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive Human-Machine Conversation with Explicit Conversation Goals. *arXiv preprint arXiv:1906.05572* (2019).

[157] Yu Wu, Zhoujun Li, Wei Wu, and Ming Zhou. 2018. Response selection with topic clues for retrieval-based chatbots. *Neurocomputing* 316 (2018), 251–261.

[158] Yu Wu, Furu Wei, Shaohan Huang, Zhoujun Li, and Ming Zhou. 2019. Response Generation by Context-aware Prototype Editing. In *Proceedings of AAAI 2019, Honolulu, Hawaii, USA, January 27-February 1, 2019.*

[159] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of ACL 2017, Vancouver, Canada, July 30-August 4, 2017.* 496–505.

[160] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of AAAI 2017, San Francisco, California, USA., February 4-9, 2017.* 3351–3357.

[161] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-Promoting GAN: A Cross-Entropy Based Generative Adversarial Network for Diversified Text Generation. In *Proceedings of EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018.* 3940–3949.

[162] Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training. *arXiv preprint arXiv:1809.04505* (2018).

[163] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of SIGIR 2016, Pisa, Italy, July 17-21, 2016.* 55–64.

[164] Rui Yan and Dongyan Zhao. 2018. Smarter Response with Proactive Suggestion: A New Generative Neural Conversation Paradigm. In *Proceedings of IJCAI 2018, Stockholm, Sweden, July 13-19, 2018.* 4525–4531.

[165] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A Hybrid Retrieval-Generation Neural Conversation Model. *arXiv preprint arXiv:1904.09068* (2019).

[166] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceedings of SIGIR 2018, Ann Arbor Micigan, USA, July 8-12, 2018.* 245–254.

[167] Min Yang, Zhou Zhao, Wei Zhao, Xiaojun Chen, Jia Zhu, Lianqiang Zhou, and Zigang Cao. 2017. Personalized Response Generation via Domain adaptation. In *Proceedings of SIGIR 2017, Tokyo, Japan, August 7-11, 2017.* 1021–1024.

[168] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018.* 4970–4977.

[169] Zhou Yu, Ziyu Xu, Alan W. Black, and Alexander I. Rudnicky. 2016. Strategy and Policy Learning for Non-Task-Oriented Conversational Systems. In *Proceedings of SIGDIAL 2016, Los Angeles, CA, USA, September 13-15, 2016.* 404–412.

[170] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018.* 1350–1361.

[171] Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding Your Voice: The Linguistic Development of Mental Health Counselors. *arXiv preprint arXiv:1906.07194* (2019).

[172] Jiayi Zhang, Chongyang Tao, Zhenjing Xu, Qiaojing Xie, Wei Chen, and Rui Yan. 2019. EnsembleGAN: Adversarial Learning for Retrieval-Generation Ensemble Model on Short-Text Conversation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 435–444.

[173] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018. Learning to Control the Specificity in Neural Response Generation. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018.* 1108–1117.

[174] Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir R. Radev. 2018. Addressee and Response Selection in Multi-Party Conversations With Speaker Interaction RNNs. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018.* 5690–5697.

[175] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018.* 2204–2213.

[176] Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2017. Neural personalized response generation as domain adaptation. *World Wide Web* (2017), 1–20.

[177] Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Consistent Dialogue Generation with Self-supervised Feature Learning. *arXiv preprint arXiv:1903.05759* (2019).

[178] Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2019. Improving Background Based Conversation with Context-aware Knowledge Pre-selection. *arXiv preprint arXiv:1906.06685* (2019).

[179] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. arXiv:cs.CL/1911.00536

[180] Zheng Zhang, Minlie Huang, Zhongzhou Zhao, Feng Ji, Haiqing Chen, and Xiaoyan Zhu. 2019. Memory-augmented Dialogue Management for Task-oriented Dialogue Systems. *ACM Transactions on Information Systems* 1 (2019).

[181] Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3740–3752.

[182] Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural Multimodal Belief Tracker with Adaptive Attention for Dialogue Systems. In *The World Wide Web Conference*. ACM, 2401–2412.

[183] Tiancheng Zhao and Maxine Eskénazi. 2016. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning. In *Proceedings of SIGDIAL 2016, Los Angeles, CA, USA, September 13-15, 2016*. 1–10.

[184] Tiancheng Zhao, Kyusong Lee, and Maxine Eskénazi. 2018. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018*. 1098–1107.

[185] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of ACL 2017, Vancouver, Canada, July 30-August 4, 2017*. 654–664.

[186] Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized Dialogue Generation with Diversified Traits. *CoRR* abs/1901.09672 (2019).

[187] Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-Aware Neural Machine for Dialogue Response Generation. In *Proceedings of AAAI 2017, San Francisco, California, USA., February 4-9, 2017*. 3400–3407.

[188] Ganbin Zhou, Ping Luo, Yijun Xiao, Fen Lin, Bo Chen, and Qing He. 2018. Elastic Responding Machine for Dialog Generation with Dynamically Mechanism Selecting. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018*. 5730–5737.

[189] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *Proceedings of AAAI 2018, New Orleans, Louisiana, USA, February 2-7, 2018*. 730–739.

[190] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of IJCAI 2018, Stockholm, Sweden, July 13-19, 2018*. 4623–4629.

[191] Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A Dataset for Document Grounded Conversations. In *Proceedings of EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018*. 708–713.

[192] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *CoRR* abs/1812.08989 (2018).

[193] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view Response Selection for Human-Computer Conversation. In *Proceedings of EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 372–381.

[194] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018*. 1118–1127.

[195] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018*. 1128–1137.

[196] Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible End-to-End Dialogue System for Knowledge Grounded Conversation. *CoRR* abs/1709.04264 (2017).

Name	Topic	Source	Language	Corpus Statistics	Corpus Features
STC[123]	Open topics	Social media (Weibo)	Chinese	219,905 posts 4,308,211 responses	One post multiple responses
Twitter Triple[129]	Open topics	Social media (Twitter)	English	29M (c,m,r) triples[1] 4,232 test/val triples	Context information
Ubuntu Dialog[79]	Ubuntu technical issues	Online chat log	English	930,000 dialogs 7.71 turns per dialog 10.34 words per turn	Task-specific dialog
PersonalDialog[186]	Open topics	Social media (Weibo)	Chinese	20.83 million dialogs 56.26M utterances 8.47M user profiles	Personalization, rich user profiles
Persona-Chat[175]	Daily life	Crowd source	English	10,981 dialogs 164,356 utterances	Personalization
DailyDialog[66]	Daily life	Web	English	13,118 dialogs 7.9 turns per dialog 14.6 words per turn	Emotion and intent annotation
CMU DOG[191]	30 movies' wikipedia page	Crowd source	English	4,112 dialogs 31.6 turns per dialog 10.8 words per turn	Knowledge-grounded
Holl-E[86]	921 movies	Crowd source	English	9,071 dialogs 10.0 turns per dialog 15.3 words per turn	Knowledge-grounded
Wizard of Wikipedia[25]	1,365 Wikipedia articles	Crowd source	English	22,311 dialogs 9.0 turns per dialog	Knowledge-grounded
Grounded Response Generation DSTC7 [104]	Web articles	Reddit	English	32.7K dialog-document pairs 2.8M utterances 17M document sentences	Knowledge-grounded
Topical-Chat[40]	8 domains, e.g. politics, fashion	Crowd source	English	11,319 dialogs 22 turns per dialog 19.8 words per turn	Knowledge-grounded
OpenDialKG[87]	Movie, book, sports, music	Crowd source	English	15,673 dialogs 91,209 turns	Knowlege-grounded
DuConv[156]	Films and film stars	Crowd source	Chinese	29,858 dialogs 9.1 turns per dialog 10.6 words per turn	Knowledge-grounded/ Proactivity modeling
DyKgChat[17]	2 TV series	TV series	Chinese English	1,247/3,092 dialogs[2] 13.8/18.7 turns per dialog[2] 27.0/16.5 words per turn[2]	Knowledge-grounded
Empathetic Dialogues[109]	Dialy life	Crowd source	English	24,850 dialogs 4.31 turns per dialog 15.2 words per turn	Emotional/empathetic dialog modeling
Target-Guided Conversation [135]	Daily life	Crowd source	English	8,939 dialogs 101,935 utterances 2,678 keywords	Proactivity, behavior and strategy
PERSUASION-FOR-GOOD [147]	Charity donation	Crowd source	English	1,017 dialogs 10.43 turns per dialog 19.36 words per utterance	Personalization, behavior and strategy

Table 3. Open-domain Dialog Corpora. We only list the datasets that are frequently used or recently proposed[3].

[1] (c,m,r) means a triple of (context, message, response).
[2] The first number is for the Chinese TV series and the second for the English one.
[3] A complete survey on older datasets was published in 2015 [118] so that we do not include those corpora.