Open-domain Dialogue Generation: What We Can Do, Can*not* Do, And Should Do Next

Katharina Kann, Abteen Ebrahimi, Joewie J. Koh, Shiran Dudy, Alessandro Roncone
University of Colorado
first.last@colorado.edu

Abstract

Human–computer conversation has long been an interest of artificial intelligence and natural language processing research. Recent years have seen a dramatic improvement in quality for both task-oriented and open-domain dialogue systems, and an increasing amount of research in the area. The goal of this work is threefold: (1) to provide an overview of recent advances in the field of open-domain dialogue, (2) to summarize issues related to ethics, bias, and fairness that the field has identified as well as typical errors of dialogue systems, and (3) to outline important future challenges. We hope that this work will be of interest to both new and experienced researchers in the area.

1 Introduction

Being an empathetic, entertaining, and knowledgeable dialogue partner can be difficult even for humans. Unsurprisingly, the task of dialogue generation, i.e., creating a system that is able to hold an intelligent conversation in a way a human would, constitutes a hard challenge for the natural language processing (NLP) community. In recent years, partially due to the development of powerful natural language understanding (NLU) and natural language generation (NLG) models (Radford et al., 2018; Devlin et al., 2019), the quality of dialogue systems has been improving.

Systems fall into two broad categories, depending on if they support task-oriented or open-domain dialogues. Task-oriented dialogue systems are built for specific purposes, such as booking a flight, and the topic of conversation is limited to the domain of interest. While a narrow scope reduces the complexity of the task, the fact that misunderstandings can have severe consequences adds to it: exact understanding of the user's intentions is crucial. In contrast, open-domain dialogue systems have the ability to talk about a wide variety of arbitrary topics. Thus, conversations with open-domain dialogue systems more closely resemble human–human conversations. Users often do not have any specific goal beyond enjoying the conversation. Over the last few years – boosted by the development of deep learning models for text – the NLP community has seen rapid advances in the area of dialogue generation. A consequence of this success, as well as of the general growth of the NLP community, has been an abundance of publications on the topic: 275 submissions made *Dialogue Systems* the fourth largest track at ACL 2021 in terms of submitted papers.[1]

Utterance	**Fluent**	**Meaningful**	**Engaging**
I have never been to Italy.	✓		
Mulan I yesterday		✓	
I saw Mulan yesterday.	✓	✓	
I saw Mulan yesterday and it was great – have you seen it?	✓	✓	✓

Table 1: Possible responses of an open-domain dialogue system to *Have you recently seen a good movie?*

To assist researchers in keeping up with the fast progress and to provide a starting point for newcomers, we aim at providing a comprehensive overview of what we as a field currently can do (*existing research*), what we yet cannot do (*common errors of dialogue systems*) or believe must not do (*problems related to ethics, bias, and fairness*), and what we should do (*open challenges for open-domain dialogue generation*). Our work complements Serban et al. (2015), Finch and Choi (2020), and Huang et al. (2020) – surveys of dialogue datasets, evaluation techniques, and model architectures, respectively, by providing a holistic view of the field.

2 Open-domain Dialogue Generation

We use the following definition for open-domain dialogue generation, the task of a *social chatbot* or *socialbot*: Given zero or more previous dialogue

[1]These numbers are based on statistics presented during the opening session of ACL-IJCNLP 2021.

turns between itself and one or more other participants, a system must output a fluent, engaging, and meaningful natural language response. Table 1 shows example outputs of low and high quality. In general, the conversation should continue until all human participants signal that it should end. An open-domain dialogue further does not have to have an explicit goal, i.e., it does not have to center around a task to solve. The conversation can further shift between topics or *domains*, e.g., from movies to politics to sports. While an ideal open-domain dialogue system would also handle task-oriented parts of the conversation, this is not yet common practice. Thus, we consider open-domain and task-oriented dialogue to be mutually exclusive for the purpose of this survey.

Task evaluation. Evaluation strategies can be sorted into two broad categories: *automatic metrics* and *human evaluation*. Automatic metrics are cheap, but do not always correlate well with human judgments (Liu et al., 2016). Common metrics for generative systems are perplexity (Vinyals and Le, 2015), BLEU (Papineni et al., 2002; Ghazvininejad et al., 2017), or DIST-n (Li et al., 2016a). For retrieval systems, recall at position k in n candidates ($R_n@k$), mean average precision (MAP), mean reciprocal rank (MRR) and precision at position 1 (P@1) are used (Wu et al., 2017).

Human evaluation is expensive, but done frequently, due to a lack of good automatic alternatives (Shang et al., 2015; Ram et al., 2018b). For instance, Deriu et al. (2020) propose to evaluate models by determining from which point in a conversation on one can tell they are not human.

A detailed description of open-domain dialogue evaluation goes beyond the scope of this paper. We refer the interested reader to a recent survey on the subject by Finch and Choi (2020).

3 Open-domain Dialogue Datasets

English datasets. The Twitter dataset (Ritter et al., 2010) consists of roughly 1.3 million Twitter conversations with 2 to 243 posts each. Sordoni et al. (2015) generalize it to the Twitter Triples Corpus, which contains context–message–response triples. The context represents previous dialogue turns, and the response is the user's reply to the message. Adiwardana et al. (2020) mine the Meena dataset, which consists of about 867 million context–reply pairs from public posts. Each context consists of all previous utterances in the conversation that a reply is participating in.

The PersonaChat dataset (Zhang et al., 2018b) consists of chats and personas which are collections of five or more sentences that describe a personality. The dataset also contains revised personas, which are rewritten versions meant to prevent models from using simple word overlap to learn a persona. The chats are dialogues between two workers who each emulate one persona. The Target Guided Conversation Dataset (Tang et al., 2019) is derived from the PersonaChat corpus and leverages keywords for transitions between turns. The persona information is removed, and a rule-based keyword extractor is used to find keywords. This dataset allows for models to proactively guide the user towards a target topic. Similar to the PersonaChat dataset, the Wizard of Wikipedia dataset (Dinan et al., 2019) consists of dialogues between two crowdworkers: now, one worker is a "wizard" and the other an "apprentice". The wizard is given text about a topic from Wikipedia, and the two are told to converse about it. The wizard labels each of their utterances with a sentence in the article that provides the knowledge used. The dataset is meant to aid creating dialogue systems that are able to use knowledge in retrieving or generating responses.

OpenDialKG (Moon et al., 2019) is created by asking two workers to converse about a topic using facts from a KG. One worker is given an entity and told to start a conversation about it. The second worker is given facts and told to respond using the most natural and relevant-sounding fact. As the conversation evolves, KG entities are surfaced to allow workers to use them in their responses. Another grounded dataset is the CMU Document Grounded Dataset (Zhou et al., 2018). The authors give workers a Wikipedia article on a movie, and ask them to converse about it for at least 12 turns. 2 experimental scenarios are considered: in the first, only one worker is given the article, and is told to convince the other person to watch it; in the second, both workers are given the article, and they are instructed to talk about the content. In a similar vein, Qin et al. (2019) create a large corpus of grounded conversations by scraping comments between users on Reddit. They consider threads where users are discussing entities found in a linked web document. Due to the common use of anchors to relevant information in the URLs of linked documents, the authors use this dataset to train systems which can take advantage of machine reading comprehension models. The Topical-Chat Corpus (Gopalakrishnan et al., 2019) is a grounded corpus built using 300 entities across 8 topics. Two workers are given reading sets, which are a collection of crowdsourced fun facts, Washington Post articles, and condensed Wikipedia lead sections. Different reading set configurations allow for a potentially asymmetrical amount of information to be given to each person. Conversations are required to have a minimum of 20 turns, and workers are asked to annotate the sentiment of their utterances, where they found the information they spoke about, and the quality of their partner's utterances. The DailyDialog dataset (Li et al., 2017b) is created by scraping text from conversations held on an English learning website. Each utterance is labeled with a dialogue act and an emotion.

The EmpatheticDialogues dataset (Rashkin et al., 2019) contains conversations grounded in situation descriptions. To get these situation descriptions, crowdworkers are asked to write about an emotional situation. Subsequently, two workers are paired up and given a situation to roleplay. The goal of the dataset is to help to train systems that can identify user emotion from dialogue text. Li et al. (2020c) also give workers roles in order to create the AntiScam dataset. It consists of dialogues between crowdworkers, where one worker is assigned the role of an attacker and the other the role of a user. In their conversations, the attacker poses as an Amazon customer service agent and attempts to collect the user's information. The Persuasion for Social Good dataset (Wang et al., 2020b) contains conversations between two crowdworkers, one of whom is trying to convince the other to donate to a specific charity. 300 of these conversations are annotated with one of ten persuasion strategies, or marked as a non-strategy. The objective of collecting this data is to improve the persuasiveness of dialogue agents.

Chinese datasets. Song et al. (2020) introduce the Key-value Profile Identification dataset (KvPI). This data comes from the Sina Weibo social network and consists of text in Mandarin Chinese. KvPI contains post–response pairs, along with three attributes describing the poster (gender, location, and constellation). Each post–response pair is annotated as either entailing, contradicting, or being irrelevant to an attribute. This dataset is designed to investigate how to automatically detect consistency between dialogue posts and the dialogue agent's profile. The Weibo dataset (Wang et al., 2013) is a standard open-domain dialogue generation corpus. Similar to the aforementioned ones it is collected from Sina Weibo. It contains about 0.6 million query–response pairs. Also from Weibo, Shang et al. (2015) create the Short Text Conversation Corpus. Utterance pairs are matching posts and their replies. PersonalDialog (Zheng et al., 2019) was also collected from Weibo. Multi-turn conversations were created by taking user posts and their comments, and each utterance is connected with a specific person, who is represented by a key-value dictionary of traits. This dataset allows to incorporate personality information into generated responses. The PChatbot dataset is collected by Qian et al. (2021) from Weibo posts and Chinese judicial forums. It is composed of almost 200 million dialogue pairs. Each utterance is linked to an anonymized user ID. One potential use for this dataset is to have a model learn to respond differently to users depending on their dialogue history.

Wu et al. (2017) present the Douban dataset, which consists of conversations between two people on the Douban social network. All but the last utterance of each conversation are considered the context and the last utterance is considered an appropriate response. The Douban dataset further contains an additional test set that consists of contexts from Douban posts paired with final utterances from the Weibo that are labeled by humans as positive or negative matches based on the context. The E-commerce dataset (Zhang et al., 2018c) consists of conversations between Chinese customers and customer service staff. As in the Douban dataset, the last utterance is considered a positive response for the rest of the conversation. Negative responses are retrieved from other conversations in an automated fashion. The E-commerce and Douban datasets can be used for training and testing retrieval-based multi-turn dialogue systems.

DuConv (Wu et al., 2019) is a KG-based dataset. A KG is created from information about movies and their characters. To create conversations, first a "conversation path" is created by finding a path between two sampled entries in the KG. Then, two crowdsource workers are given roles – leader and follower – and asked to converse. The leader has access to the conversation path and the KG, and the follower only has access to the leader's utterances. The conversation continues until the leader reaches the conversation goal. DyKgChat (Tuan et al., 2019) was created by scraping conversations from two TV shows, one in Chinese, and one in English. Additionally, manually created KGs are provided to cover entities from the shows.

Finally, Chen and Kan (2013) collect NUS SMS, consisting of over 70,000 SMS messages in both Chinese and English.

Multilingual and multimodal datasets. Open-domain dialogue datasets in languages besides English and Chinese are difficult to find. A Korean dataset has been created by Kim et al. (2021) by translating the English Wizard of Wikipedia dataset (Dinan et al., 2019). To the best of our knowledge, the only *multilingual* dataset is XPersona (Lin et al., 2020a), an extension of the English PersonaChat dataset (Zhang et al., 2018b) to Chinese, French, Indonesian, Italian, Korean, and Japanese. It is created by first automatically translating the training, development and test data. The latter two splits are then manually corrected, while the training set only receives semi-manual cleaning. The authors use this dataset to evaluate approaches based on multilingual models and automatic translation.

Multimodal datasets also exist: Image-Chat by Shuster et al. (2020) consists of images together with English dialogues. Each dialogue is linked to a pair of styles or emotions portrayed in the dialogue. The images are of everyday things, such as food or landscapes. The dialogues are from conversations between two crowd workers who are asked to discuss the image and each given a style or emotion to portray in their discussion. This dataset aims at creating dialogue systems that can speak in different styles and express varying emotions. Meng et al. (2020) present OpenViDial, which consists of dialogues and their visual contexts from movies and TV series. MMChat (Zheng et al., 2021a) contains Chinese conversations about images, which have been scraped from Weibo.

We refer interested readers to Serban et al. (2015) for more information on corpora; for a table with all datasets mentioned here see Appendix A.

4 Open-domain Dialogue Systems

We sort approaches into three categories: (1) *retrieval systems*, which get their responses from a dataset; (2) *generative systems*, which generate responses automatically; and (3) *comprehensive systems*, which consist of a dialogue manager (DM), at least one system from the aforementioned categories, and optionally other functional modules.

4.1 Retrieval Systems

Retrieval systems first obtain a candidate response set from a large repertoire of options and then determine how well each candidate suits the dialogue context. Models can be arbitrarily complex and operate on a single-turn (Wang et al., 2013) or multi-turn (Wu et al., 2017) basis. As retrieval systems do not have a generative component and their outputs originate from human conversations, they are generally fluent and understandable. They are also relatively safe, as many types of harmful responses can be filtered. However, retrieval systems are limited in their ability to converse about topics not covered in the provided responses.

Non-neural approaches exist, such as support-vector machine (SVM)-based ones (Wang et al., 2013; Ji et al., 2014). More recently, neural models which compute the matching score between candidate responses and dialogue contexts have been developed. Initially, feed-forward networks have been employed (Lu and Li, 2013). Wang et al. (2015) extend prior approaches by representing both a candidate response and the context as dependency trees and extracting features from those representations, before obtaining their score via a deep feed-forward network. Later work has used a combination of convolutional neural network (CNN) and recurrent neural network (RNN) layers to determine the matching scores of possible responses, sometimes in combination with an attention mechanism (Yan et al., 2016; Zhou et al., 2016; Wu et al., 2017; Zhang et al., 2018c; Tao et al., 2019). Lu et al. (2019) add spatio-temporal features to their model. The multi-hop selector network by Yuan et al. (2019) looks for the relevant context in a multi-turn dialogue, and uses the context utterances determined to be relevant when retrieving a response. The dually interactive matching network (Gu et al., 2019b) retrieves responses based on personas. It extends Li et al. (2016b) to the previously proposed interactive matching network (Gu et al., 2019a).

Retrieval systems can also be based on transformers (Vaswani et al., 2017). The transformer memory network, for instance, takes knowledge from the Wizard of Wikipedia dataset to retrieve more knowledge-focused responses (Dinan et al., 2019). Whang et al. (2020) go one step further and use *pretrained* transformer models, namely BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020), for matching. With this, they follow earlier work on response retrieval for domain-specific dialogue systems. They further add multi-task training. Gao et al. (2020) propose a DialoGPT (Zhang et al., 2020c)-based model to rank retrieved responses.

Lin et al. (2020b) propose to train retrieval models using a ranking loss and so-called grey-scale data: they construct training examples from ground-truth, generated, and random responses.

4.2 Generative Systems

Generative systems generate responses freely, i.e., they are not limited to a predefined set of utterances. Their responses are not guaranteed to be well-formed. However, in contrast to retrieval systems, they are not restricted to talking about topics within a predefined set of responses.

The arguably first generative dialogue system has been ELIZA (Weizenbaum, 1966). ELIZA is rule-based and plays the role of a therapist. Parry, in contrast, is designed to act like a psychology *patient* (Colby, 1975). Later, ALICE has been created by Wallace (1995) as a proof of concept for the Artificial Intelligence Markup Language.

The large majority of generative systems are neural sequence-to-sequence (seq2seq) models. The first such models have been created by Shang et al. (2015) and, concurrently, Vinyals and Le (2015). Their systems are LSTM-based seq2seq models. Parthasarathi and Pineau (2018) add two knowledge sources to an LSTM seq2seq model: the NELL knowledge base (Carlson et al., 2010) and Wikipedia summaries (Scheepers, 2017). Li et al. (2016b) propose a persona-based LSTM encoder-decoder. They represent personas via sentences, with a persona vector being the combination of the sentences. Similarly, Zhang et al. (2018b) condition a dialogue system on profile sentences and also build profiles of its users, allowing it to better tailor its responses to individuals.

Luo et al. (2018)'s LSTM seq2seq model is able to learn utterance-level semantic dependencies, which makes responses more coherent and fluent. Furthermore, Li et al. (2020b) propose two additions to a standard LSTM model: a rank-aware calibrator network, used to construct contrastive optimization objectives, and a knowledge inference component, which learns keywords in order to help the model use more informative words during generation. Zhang et al. (2020a) use a GRU-based response generation model along with a deep utterance aggregation model to generate a context vector from previous turns.

Ghazvininejad et al. (2017) leverage a facts dataset to inject knowledge into a GRU seq2seq model, which helps the model generate more knowledgeable responses. A collection of synonym sets was used by Hsueh and Ma (2020) to help address the problem of social chatbots repeatedly responding with similarly worded sentences.

A variational hierarchical recurrent encoder-decoder (VHRED) for open-domain dialogue generation is proposed by Serban et al. (2017). This model uses latent stochastic variables to model hierarchical structure between dialogue turns, and feeds that information into an RNN. Subsequently, Zhao and Kawahara (2020) introduce a VHRED with a linear Gaussian prior.

Transformer-based models include generative variants of the transformer memory network (Dinan et al., 2019). Further, Keskar et al. (2019) train a conditional transformer language model, which accepts various control codes as part of the input. These control codes allow the control of style, content, and other behaviors without requiring the model to be retrained. Meena (Adiwardana et al., 2020) is a transformer-based seq2seq model trained on large amounts of real chat data. Know-EDG (Li et al., 2020a) consists of a knowledge-enhanced context encoder and an emotion identifier linear layer in front of a transformer model. The input from the emotion identifier allows the model to alter its generated responses based on the emotion its dialogue partner is expressing. Zheng et al. (2021b) add style embeddings to a transformer-based system to alter its dialogue style. Dziri et al. (2021) tackle the problem of factually untrue responses with a generate-then-refine strategy: generated responses are corrected with the help of a knowledge graph.

A mixture between a retrieval and a generative system is the RetrieveNRefine model (Weston et al., 2018). It first employs a key-value memory network to retrieve a good dialogue response, which is then refined by an LSTM seq2seq model.

Only recently, multimodal dialogue models, which combine language and image processing components have been developed (Shuster et al., 2020). Shuster et al. (2021) explore the integration of large pretrained transformer models for text into such systems.

4.3 Comprehensive Systems

Comprehensive systems consist of multiple components together with a DM. They are typically not trained in an end-to-end fashion. The DM selects one or more of the available – in some cases highly specialized – response generators to produce a response for a given context.

XiaoIce (Zhou et al., 2020) is a comprehensive system which consists of 3 layers: The user experience layer connects the system to social media and chat services. The conversation engine layer contains a core chat module, a skills module, a DM, and an empathetic computing module. Finally, the data layer contains profile information on XiaoIce and users, knowledge graphs (KGs), topic indices, and other information. Adapter-Bot (Madotto et al., 2020) employs a DM which is based on BERT (Devlin et al., 2019), a backbone conversational model based on DialoGPT (Zhang et al., 2020c), and a series of additional smaller modules.

Alexa Prize competition. The Amazon Alexa Prize (AP) is an annual competition, with the grand challenge of designing a system capable of holding an open-domain conversation for 20 minutes (Ram et al., 2018a). Contestants develop live systems which are randomly selected to converse with Alexa users. Once the conversation is finished, users are requested to give a rating, which is the main metric used for evaluation. The teams with the highest rating move on to the finals, where expert judges decide the winner.

Sounding Board (Fang et al., 2017), which won the inaugural AP in 2017, is a comprehensive dialogue system which is comprised of an NLU module, a DM, topic-specific modules with rule-based mini-skills, and an NLG component. The NLU module uses a series of text classifiers to extract the user's primary intent. The DM receives that information and, using a hierarchical rule-based architecture, decides which of the mini-skills to use when generating dialogue acts and content to pass to the NLG module. The NLG module builds a response in a rule-based fashion. Gunrock (Chen et al., 2018), the winner of the 2018 AP, differs from Sounding Board in the techniques used for each piece. The NLU module contains multiple submodules, including a noun phrase extractor, a topic model, and a sentiment analyzer. The information from these submodules is passed to the DM, which selects a topic and activates the corresponding submodules. The information from the NLU module and the topic submodule is then passed to the NLG module, which builds a response using templates. Gunrock 2.0 entered the 2019 AP (Liang et al., 2020), and differs from its predecessor by relying more on neural models. However, the 2019 AP was won by Emora (Finch et al., 2020). In addition to mentioning facts, Emora also supports talking about experiences and opinions. Besides the winning system, finalists of the 2019 AP include Chirpy Cardinal (Paranjape et al., 2020), which employs generators based on GPT-2 (Radford et al., 2019), and Alquist (Pichl et al., 2020), which relies on conversation graphs to dynamically use knowledge in its responses. Many design choices were common among other contenders. For NLU, systems often use dialogue act, topic, and intent classifiers. Systems also rely heavily on named entity recognition and entity linking, such as Tartan (Chen et al., 2020), whose response generators use a knowledge base for slot filling. Other systems employ a mixture of strategies to generate responses, such as Athena (Harrison et al., 2020), which attempts to switch between rule-based, knowledge-based, and retrieval-based modules on-the-fly, as well as DREAM (Kuratov et al., 2020), which employs candidate and response annotators before serving a final response. Other contenders include Audrey (Hong et al., 2020), which focuses on emotion and personality, Zotbot (Schallock et al., 2020), which incorporates a commonsense-reasoning element, and Bernard (Majumder et al., 2020), which is built around non-deterministic finite automata.

5 Training and Data Augmentation

Retrieval-based systems are commonly trained with a cross-entropy loss (Zhang et al., 2018c; Lu et al., 2019), comparing a prediction against the gold standard from a training set. As an alternative, using a ranking loss, where a model is trained on distinguishing suitable from unsuitable responses, has been proposed (Lin et al., 2020b). In *comprehensive systems*, the individual components are usually trained separately.

Several algorithms to train *generative systems* have been proposed. Given a training set $D = \{(R_1, C_1, B_1), \ldots, (R_N, C_N, B_N)\}$ with N examples consisting of context C_i, background information B_i, and response R_i, models are most commonly trained using maximum likelihood estimation (Shang et al., 2015; Vinyals and Le, 2015).

The goal is to minimize the loss

$$L = -\sum_{i=1}^{N} \log P(R_i | C_i, B_i). \tag{1}$$

However, it has been shown that this encourages boring responses (Li et al., 2016a). As a remedy, several ways to weight training examples have been proposed (Shang et al., 2018; Li et al., 2020b). With that, the loss changes to

$$L = -\sum_{i=1}^{N} w_i \log P(R_i | C_i, B_i), \tag{2}$$

where w_i is the weight corresponding to example i. Further, Zhao and Kawahara (2020) address the concern that generally multiple responses are possible. They propose multi-referenced training and automatically create M different responses $\widetilde{R}_{im}$ for each original R_i. Their loss is

$$L = -\sum_{i=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \log P(\widetilde{R}_{im} | C_i, B_i). \tag{3}$$

Contrastive learning (Hadsell et al., 2006; Gutmann and Hyvärinen, 2012; Cai et al., 2020a) – where a model is trained to assign higher and, respectively, lower conditional probabilities to positive and negative samples than a reference model – and curriculum learning – during which examples are presented to a model in a specific order – have also been employed (Cai et al., 2020c). Finally, dialogue systems can also be trained via reinforcement learning (Li et al., 2016c; Zhang et al., 2018a; Sankar and Ravi, 2019) or adversarial learning (Li et al., 2017a).

Pretraining. Large pretrained models such as BERT (Devlin et al., 2019) or GPT and its successors (Radford et al., 2018, 2019; Brown et al., 2020) have improved the state of the art for a variety of NLP tasks. Pretraining has also been used for open-domain dialogue generation. Two different strategies exist: One option is to pretrain a model on large unlabeled corpora to then finetune it on dialogue data. Liu et al. (2020c), for instance, initialize parts of their generative system with a pretrained BERT model, and Gu et al. (2020) finetune BERT for multi-turn response selection in retrieval-based chatbots. Shi et al. (2020) introduce an English language-learning chatbot based on GPT-2. Boyd et al. (2020) condition a GPT-2 model for dialogue generation on several previous conversations of a single individual to get it to use that individual's style. Further, plug and play language models consist of pretrained language models in combination with one or more simple attribute classifiers, which control various aspects of its behavior, such as style or dialogue content (Dathathri et al., 2020).

The second option is to pretrain a model on large dialogue corpora, such that it can then be finetuned on out-of-domain dialogue data. DialoGPT (Zhang et al., 2020c) is such a model. Its architecture resembles GPT, i.e., it is a transformer (Vaswani et al., 2017) language model. For training, specifically collected Reddit data is used. Like GPT, DialoGPT is publicly available. The authors also experiment with GPT-2 as a basis for DialoGPT and, similar to the work mentioned in the last paragraph, find pretraining on raw text to be beneficial. ConveRT (Henderson et al., 2020) is another model which is pretrained on dialogue data: pretraining is done on a response selection task using Reddit.

Data augmentation. Data augmentation, i.e., the creation of artificial training examples, can help in the low-resource setting. Zhang et al. (2020b) augment paired dialogue data using unlabeled data in the form of unpaired dialogue data. A dialogue pair consists of a social media post and a corresponding response. Their method starts by randomly selecting a sentence from the unpaired dataset. Then, posts that are semantically similar to the randomly selected sentence are retrieved from the paired dataset. Next, responses corresponding to the posts are collected from the paired dataset. Finally, sentences that are semantically similar to the responses are pulled from the unpaired data. Each of these newly pulled sentences are matched with the original randomly selected sentence, to create a set of candidate pairs. Those candidate pairs are then ranked, and the top-ranked pairs are saved for later use.

Other approaches differ from the aforementioned in that they do not require unlabeled data. Li et al. (2019) propose a conditional variational autoencoder as a generative data augmentation model. They combine this with a discriminator, which decides whether the generated responses are suitable for a given query. Cai et al. (2020b) design a data augmentation and instance weighting model which is trained using gradient descent and the model's performance on development examples.

6 Common Errors of Dialogue Systems

We now discuss errors common across multiple systems, considering mistakes at the turn level, the conversation level, and the system level.

Turn level. At the turn level, errors consist mostly of system responses being either *ungrammatical* or *nonsensical*. Both types of problems are more common in generative systems, as those commit errors seen in other NLG tasks, such as highly repetitive, nonsensical, or insignificant replies (Li et al., 2016a; See and Manning, 2021). Models which are motivated by semantic similarity may resort to constantly echoing the user, rather than returning a coherent response (Ritter et al., 2011; Fedorenko et al., 2018).

Conversation level. Problems arising at the conversation level are arguably more substantial than those at the turn level. Potential solutions will most likely rely heavily on advancements in other areas of NLP, such as reasoning and information extraction. A common issue consists of replies being fluent, but either not relevant in the overall context of the conversation or too generic (Adiwardana et al., 2020). Off-topic replies can often be attributed to a failure to recognize entities or previous dialogue acts. Another common problem are answers that are inconsistent across turns (Nie et al., 2021).

System level. At the system level, researchers and model developers face the difficulty of incorporating world knowledge and common sense into models (Wang et al., 2020a), as models still frequently generate responses that are factually incorrect (Mielke et al., 2020; Santhanam et al., 2021). There exists a trade-off between the range of topics a system can cover and the depth of knowledge it can leverage for any individual topic. Currently, especially comprehensive systems frequently rely heavily on curated content and static, handwritten conversation paths to talk intelligently and deeply about specific topics. However, the more a system relies on handwritten paths, the more brittle it becomes. Similarly, curated content is impossible to scale to a truly open-domain setting. Conversely, leaning more towards dynamically structured conversations gives models more flexibility and allows them to cover a wider range of topics, but often results in less meaningful responses.

7 Ethics, Bias, and Fairness

The NLP research community is becoming increasingly aware of the ethical challenges around the systems we are building, and the area of dialogue generation is no exception to this. We now summarize prior work around safety and unwanted biases.

Safety. Dialogue systems should avoid being unintentionally offensive or harming the user (Henderson et al., 2018). Therefore, attempts have been made to detect sensitive language around religion, race, violence, or contentious news as well as profanity (Tripathi et al., 2019). However, how to respond when sensitive topics are being identified is still an open question. As some of these topics shape our identities and our lives, an ideal system might not completely avoid them, and the best response strategy depends on the objectives of the system. When GPT-3 (Brown et al., 2020) and Blender (Roller et al., 2021) detect toxic language in a user utterance, they stop producing output (Xu et al., 2020). While this is an ad-hoc solution, in the long term, a graceful reaction could potentially carry the conversation to healthier places as shown by Wright et al. (2017).

Dinan et al. (2021) identify three potentially dangerous behaviors a dialogue system can exhibit: First, it can act as an *instigator* and provoke the user using negative language, as has infamously happened with the Microsoft Tay chatbot. Second, even if a system exclusively uses non-harmful language, it can cause harm to the user by being a so-called *yea-sayer*, i.e., by being overly eager to agree with the user on wrong or inappropriate statements (Lee et al., 2019; Baheti et al., 2021). Third, a dialogue system can unintentionally *impose as an expert* and provide harmful advice.

Biases. An abundance of recent work has shown that NLP models are learning undesirable biases from the data they are being trained on (Bolukbasi et al., 2016; Bordia and Bowman, 2019; Bartl et al., 2020; Shah et al., 2020). Dialogue systems are no exception to this: Liu et al. (2020a) investigate fairness in dialogue models and find that dialogue models exhibit significant prejudice against some genders and races. They propose two debiasing methods based on data augmentation and word embeddings regularization. Dinan et al. (2020b) point out that there are three types of gender bias in chat bots: the first one being due to the gender of the person that speakers are talking about, the second being due to the gender of the speaker, and the last being due to the gender of the addressee. Liu et al. (2020b) aim at mitigating the former via adversarial learning. Similarly, Dinan et al. (2020a) propose to reduce gender bias via data augmentation, targeted data collection, and bias-controlled training.

Barikeri et al. (2021) introduce RedditBias, a dataset grounded in conversations from Reddit, which enables the measurement and mitigation of gender, race, religion, and queerness bias, and use it to explore DialoGPT with and without debiasing.

8 Open Challenges for Future Research

Model evaluation and analysis. Surveying research on open-domain dialogue generation (cf. Section 4) as well as research on system evaluation (Finch and Choi, 2020), it is clear that a good automatic metric (or even manual evaluation strategy) has not yet been found. What the field needs are metrics that (1) evaluate different aspects of dialogue systems (cf. Table 1), (2) do not require references, since no reasonable set of references can contain all possibly suitable responses, and (3) correlate strongly with human judgments. One possible way to move the field towards the development of new evaluation strategies could be the establishment of a shared task on open-domain dialogue generation metrics, similar to the WMT metrics shared task (Ma et al., 2019).

Furthermore, while entire surveys are necessary to summarize work on the analysis of BERT (Rogers et al., 2020), we still know little about what dialogue systems, including DialoGPT (Zhang et al., 2020c), learn from their training data. Prior work on the analysis of dialogue models (with the exception of still non-exhaustive investigations of their biases) is limited; e.g., Saleh et al. (2020). We argue that learning more about dialogue models, which are likely to directly interact with users, is crucial. We should investigate the following: (1) What world knowledge do models acquire during training? (2) What linguistic knowledge do dialogue models learn? (3) Which potentially harmful biases do models learn from real-world data?

Multi-party dialogue. How to extend systems to handle multi-party dialogue, as posed by Seering et al. (2019), remains an underexplored area of research. Having such systems will potentially contribute to creating richer social interactions in both online and offline communities. It will further increase our understanding of the dynamics behind turn taking (Bohus and Horvitz, 2011).

Multilingual dialogue. Section 3 makes it obvious that open-domain dialogue datasets mostly exist for two high-resource languages: English and Chinese. Work on other languages is limited (e.g., Lin et al. (2020a)). We argue that, in order to speed up research on other languages, the field needs to develop datasets with the following properties: (1) datasets should be created for a diverse set of potentially low-resource languages and (2) the created datasets should not be translations of existing datasets. The latter is necessary since it has been shown for other NLP tasks that translated datasets show different properties from those natively collected in a language (Artetxe et al., 2020).

9 Conclusion

Recent years have seen a drastic improvement in the quality of open-domain dialogue systems as well as in the amount of research in the area. Therefore, we first presented an overview of the state of the field of NLP for open-domain dialogue. Then, we outlined important future challenges: better model evaluation and analysis, multi-party dialogue, and multilingual dialogue.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. We are also grateful to the members of CU Boulder's NALA group for their feedback on and input to this paper. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF.

References

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*.

Rami Al-Rfou, M. Pickett, Javier Snaider, Yun-Hsuan Sung, B. Strope, and R. Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *ArXiv*, abs/1606.00372.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Dan Bohus and Eric Horvitz. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*, pages 98–109.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Boyd, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Large scale multi-actor generative dialog modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 66–84, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hengyi Cai, Hongshen Chen, Yonghao Song, Zhuoye Ding, Yongjun Bao, Weipeng Yan, and Xiaofang Zhao. 2020a. Group-wise contrastive learning for neural dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 793–802, Online. Association for Computational Linguistics.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020b. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online. Association for Computational Linguistics.

Hengyi Cai, Hongshen Chen, Cheng Zhang, Yonghao Song, Xiaofang Zhao, Yangxi Li, Dongsheng Duan, and Dawei Yin. 2020c. Learning from easy to complex: Adaptive multi-curricula learning for neural dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7472–7479.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 1306–1313.

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*.

Fanglin Chen, Ta-Chung Chi, Shiyang Lyu, Jianchen Gong, Tanmay Parekh, Rishabh Joshi, Anant Kaushik, and Alexander Rudnicky. 2020. Tartan: A two-tiered dialog framework for multi-domain social chitchat. *Alexa prize proceedings*.

Tao Chen and Min-Yen Kan. 2013. Creating a live, public short message service corpus: the nus sms corpus. *Language Resources and Evaluation*, 47(2):299–335.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Kenneth Mark Colby. 1975. *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. Elsevier Science Inc.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in E2E conversational AI: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Denis Fedorenko, Nikita Smetanin, and Artem Rodichev. 2018. Avoiding echo-responses in a retrieval-based conversation system. In *Conference on Artificial Intelligence and Natural Language*, pages 91–97. Springer.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Sarah E. Finch, James D. Finch, Ali Ahmadvand, Ingyu, Choi, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, and Jinho D. Choi. 2020. Emora: An inquisitive social chatbot who cares for you.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Karthik Gopalakrishnan, Behnam Hedayatnia, Q. Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, R. Gabriel, and D. Hakkani-Tur. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019a. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2321–2324, New York, NY, USA. Association for Computing Machinery.

Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019b. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1845–1854, Hong Kong, China. Association for Computational Linguistics.

Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(2).

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Vrindavan Harrison, Juraj Juraska, Wen Cui, Lena Reed, Kevin K. Bowden, Jiaqi Wu, Brian Schwarzmann, Abteen Ebrahimi, Rishi Rajasekaran, Nikhil Varghese, Max Wechsler-Azen, Steve Whittaker, Jeffrey Flanigan, and Marilyn Walker. 2020. Athena: Constructing dialogues dynamically with discourse constraints. *Alexa prize proceedings*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.

Chung Hoon Hong, Yuan Liang, Sagnik Sinha Roy, Arushi Jain, Vihang Agarwal, Ryan Draves, Zhizhou Zhou, William Chen, Yujian Liu, Martha Miracky, Lily Ge, Nikola Banovic, and David Jurgens. 2020. Audrey: A personalized open-domain conversational bot. *Alexa prize proceedings*.

Cheng-Hsun Hsueh and Wei-Yun Ma. 2020. Semantic guidance of dialogue generation with reinforcement learning. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–9, 1st virtual meeting. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3).

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

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

San Kim, Jin Yea Jang, Minyoung Jung, and Saim Shin. 2021. A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 352–365, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuri Kuratov, Idris Yusupov, Dilyara Baymurzina, Denis Kuznetsov, Daniil Cherniavskii, Alexander Dmitrievskiy, Elena Ermakova, Fedor Ignatov, Dmitry Karpov, Daniel Kornev, The Anh Le, Pavel Pugin, and Mikhail Burtsev. 2020. Dream technical report for the alexa prize 2019. *Alexa prize proceedings*.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Juntao Li, Lisong Qiu, Bo Tang, Dongmin Chen, Dongyan Zhao, and Rui Yan. 2019. Insufficient data can also rock! learning to converse using smaller data with augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6698–6705.

Qintong Li, Piji Li, Zhumin Chen, and Zhaochun Ren. 2020a. Empathetic dialogue generation via knowledge enhancing and emotion dependency modeling.

Xin Li, Piji Li, Yan Wang, Xiaojiang Liu, and Wai Lam. 2020b. Enhancing dialogue generation via multi-level contrastive learning. *arXiv preprint arXiv:2009.09147*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2020c. End-to-end trainable non-collaborative dialog system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8293–8302.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, and Zhou Yu. 2020. Gunrock 2.0: A user adaptive social conversational system. *Alexa prize proceedings*.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020a. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

Zibo Lin, Deng Cai, Yan Wang, Xiaojiang Liu, Haitao Zheng, and Shuming Shi. 2020b. The world is not binary: Learning to rank with grayscale data for dialogue response selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9220–9229, Online. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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 the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020a. Does gender matter? Towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020b. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020c. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Constructing interpretive spatio-temporal features for multi-turn responses selection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 44–50.

Zhengdong Lu and Hang Li. 2013. A deep architecture for matching short texts. In *Advances in neural information processing systems*, pages 1367–1375.

Liangchen Luo, Jingjing Xu, Junyang Lin, Qi Zeng, and Xu Sun. 2018. An auto-encoder matching model for learning utterance-level semantic dependency in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 702–707, Brussels, Belgium. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2020. The adapter-bot: All-in-one controllable conversational model.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, Huanru Henry Mao, Sophia Sun, and Julian McAuley. 2020. Bernard: A stateful neural open-domain socialbot. *Alexa prize proceedings*.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.

Sabrina J Mielke, Arthur Szlam, Y-Lan Boureau, and Emily Dinan. 2020. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *arXiv preprint arXiv:2012.14983*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, and Christopher D. Manning. 2020. Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. *arXiv preprint arXiv:1809.05524*.

Jan Pichl, Petr Marek, Jakub Konrad, Petr Lorenc, Van Duy Ta, and Jan Sedivy. 2020. Alquist 3.0: Alexa prize bot using conversational knowledge graph. *Alexa prize proceedings*.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2470–2477, New York, NY, USA. Association for Computing Machinery.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, W. Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. *ArXiv*, abs/1906.02738.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

A. Ram, Rohit Prasad, C. Khatri, Anu Venkatesh, R. Gabriel, Q. Liu, J. Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, E. King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018a. Conversational ai: The science behind the alexa prize. *ArXiv*, abs/1801.03604.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018b. Conversational AI: The science behind the Alexa prize. *arXiv preprint arXiv:1801.03604*.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Abdelrhman Saleh, Tovly Deutsch, Stephen Casper, Yonatan Belinkov, and Stuart Shieber. 2020. Probing neural dialog models for conversational understanding. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 132–143, Online. Association for Computational Linguistics.

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. *arXiv preprint arXiv:1907.02848*.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*.

William Schallock, Daniel Agress, Yao Du, Dheeru Dua, Lyuyang Hu, Yoshitomo Matsubara, and Sameer Singh. 2020. Zotbot: Using reading comprehension and commonsense reasoning in conversational agents. *Alexa prize proceedings*.

Thijs Scheepers. 2017. *Improving the compositionality of word embeddings*. Ph.D. thesis, Master's thesis, Universiteit van Amsterdam.

Abigail See and Christopher Manning. 2021. Understanding and predicting user dissatisfaction in a neural generative chatbot. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12, Singapore and Online. Association for Computational Linguistics.

Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

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*.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Mingyue Shang, Zhenxin Fu, Nanyun Peng, Yansong Feng, Dongyan Zhao, and Rui Yan. 2018. Learning to converse with noisy data: Generation with calibration. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4338–4344. International Joint Conferences on Artificial Intelligence Organization.

Nuobei Shi, Qin Zeng, and Raymond Lee. 2020. The design and implementation of language learning chatbot with xai using ontology and transfer learning.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2021. Multi-modal open-domain dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4863–4883, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020. Profile consistency identification for open-domain dialogue agents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6651–6662, Online. Association for Computational Linguistics.

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 *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–11, Florence, Italy. Association for Computational Linguistics.

Rahul Tripathi, Balaji Dhamodharaswamy, Srinivasan Jagannathan, and Abhishek Nandi. 2019. Detecting sensitive content in spoken language. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 374–381. IEEE.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Richard Wallace. 1995. Alice - artificial linguistic internet computer entity. https://web.archive.org/web/20171227223848/http://www.alicebot.org/.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA. Association for Computational Linguistics.

J. Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020a. Improving knowledge-aware dialogue generation via knowledge base question answering. In *AAAI*.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. *arXiv preprint arXiv:1503.02427*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2020b. Persuasion for good: Towards a personalized persuasive dialogue system for social good.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong hun Lee, and Saebyeok Lee. 2020. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62, Vancouver, BC, Canada. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

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 the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

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 the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 111–120.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018a. Reinforcing coherence for sequence to sequence model in dialogue generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4567–4573. International Joint Conferences on Artificial Intelligence Organization.

Hainan Zhang, Yanyan Lan, Liang Pang, Hongshen Chen, Zhuoye Ding, and Dawei Yin. 2020a. Modeling topical relevance for multi-turn dialogue generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3737–3743. International Joint Conferences on Artificial Intelligence Organization.

Rongsheng Zhang, Yinhe Zheng, Jianzhi Shao, Xiaoxi Mao, Yadong Xi, and Minlie Huang. 2020b. Dialogue distillation: Open-domain dialogue augmentation using unpaired data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3449–3460, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018c. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tianyu Zhao and Tatsuya Kawahara. 2020. Multi-referenced training for dialogue response generation. *arXiv preprint arXiv:2009.07117*.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Ke Lin. 2021a. MMChat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*.

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021b. Stylized dialogue response generation using stylized unpaired texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14558–14567.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

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 the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Austin, Texas. Association for Computational Linguistics.

A Overview of Existing Datasets

Dataset Name	Paper	Language	Method	Source
KvPI	Song et al. (2020)	zh	Scraped	Weibo
PChatbot	Qian et al. (2021)	zh	Scraped	Weibo, Judicial
Douban	Wu et al. (2017)	zh	Scraped	Douban, Weibo
E-commerce	Zhang et al. (2018c)	zh	Scraped	Taobao
Weibo	Wang et al. (2013)	zh	Scraped	Weibo
PersonalDialog	Zheng et al. (2019)	zh	Scraped	Weibo
DuConv	Wu et al. (2019)	zh	Human-Human	-
Short Text Conversation	Shang et al. (2015)	zh	Scraped	Weibo
Switchboard	Godfrey et al. (1992)	en	Human-Human	-
Twitter Dataset	Ritter et al. (2010)	en	Scraped	Twitter
Twitter Triples	Sordoni et al. (2015)	en	Scraped	Twitter
Reddit Dataset	Al-Rfou et al. (2016)	en	Scraped	Reddit
PersonaChat	Zhang et al. (2018b)	en	Human-Human	-
Wizard of Wikipedia	Dinan et al. (2019)	en	Human-Human	-
EmphateticDialogues	Rashkin et al. (2019)	en	Human-Human	-
Meena	Adiwardana et al. (2020)	en	Scraped	Social Media
AntiScam	Li et al. (2020c)	en	Human-Human	-
Dailydialogue	Li et al. (2017b)	en	Scraped	-
Persuasion for Social Good	Wang et al. (2020b)	en	Human-Human	-
CMU Document Grounded Dataset	Zhou et al. (2018)	en	Human-Human	-
Grounded Conversation Dataset	Qin et al. (2019)	en	Scraped	Reddit
Topical Chats	Gopalakrishnan et al. (2019)	en	Human-Human	-
OpenDialKG	Moon et al. (2019)	en	Human-Human	-
Target Guided Conversation Dataset	Tang et al. (2019)	en	Human-Human	-
Image-Chat	Shuster et al. (2020)	en	Human-Human	-
OpenViDial	Meng et al. (2020)	en	Scraped	Movies/TV
MMChat	Zheng et al. (2021a)	en	Sraped	Weibo
NUS SMS	Chen and Kan (2013)	en,zh	Human-Human	SMS
Korean Wizard of Wikipedia	Kim et al. (2021)	ko	MT Human-Human	-
XPersona	Lin et al. (2020a)	zh,fr,ind,it,ko,ja	MT Human-Human	-

Table 2: Overview of existing dialogue datasets. *Human–Human* denotes datasets where two people converse with each other. *Scraped* marks datasets which are gathered from an existing online resource.