Open-domain Dialogue Generation:
What We Can Do, Cannot 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.

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

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Fluent Meaningful Engaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have never been to Italy.</td>
<td>✓</td>
</tr>
<tr>
<td>Mulan I yesterday</td>
<td>✓</td>
</tr>
<tr>
<td>I saw Mulan yesterday</td>
<td>✓</td>
</tr>
<tr>
<td>I saw Mulan yesterday and it was great – have you seen it?</td>
<td>✓</td>
</tr>
</tbody>
</table>

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

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

¹These 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 crowd-workers, 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 crowdsourced 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 RetrieveNERefine 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.

Xiaolce (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 Xiaolce 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, Zobot (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

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

(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

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

(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 \( \tilde{R}_{im} \) for each original \( R_i \). Their loss is

\[ L = -\sum_{i=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \log P(\tilde{R}_{im}|C_i, B_i). \]  

(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. DialogGPT (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, DialogGPT is publicly available. The authors also experiment with GPT-2 as a basis for DialogGPT 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) augmented 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 chatbots: 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


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.


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.


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


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


Rahul Tripathi, Balaji Dhamodharaswamy, Srinivasan Jaganmathan, 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.


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

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.


Shi 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.


Zhuosheng Zhang, Jianqiang Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018c. Modeling multi-turn con-


### A Overview of Existing Datasets

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

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.