Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges

Shikib Mehri[1], Jinho Choi[2], Luis Fernando D'Haro[3], Jan Deriu[4], Maxine Eskenazi[1], Milica Gasic[5], Kallirroi Georgila[6], Dilek Hakkani-Tur[7], Zekang Li[8], Verena Rieser[9], Samira Shaikh[10], David Traum[6], Yi-Ting Yeh[1], Zhou Yu[11], Yizhe Zhang[12], Chen Zhang[13]

[1]**Carnegie Mellon University,**
[2]**Emory University,**
[3]**Universidad Politécnica de Madrid,**
[4]**Zurich University of Applied Sciences,**
[5]**Heinrich Heine University Dusseldorf,**
[6]**University of Southern California,**
[7]**Amazon Alexa AI,**
[8]**University of Chinese Academy of Sciences,**
[9]**Heriot-Watt University,**
[10]**University of North Carolina at Charlotte,**
[11]**Columbia University,**
[12]**Microsoft Research,**
[13]**National University of Singapore**

Abstract

This is a report on the *NSF Future Directions Workshop on Automatic Evaluation of Dialog*. The workshop explored the current state of the art along with its limitations and suggested promising directions for future work in this important and very rapidly changing area of research.

1. Introduction

The *NSF Future Directions Workshop on Automatic Evaluation of Dialog*[1] took place virtually on July 7 - 8, 2021. This workshop was organized by Shikib Mehri, Yi-Ting Yeh and Maxine Eskenazi of Carnegie Mellon University. The participants in the workshop have extensive experience in dialog assessment. The workshop consisted of discussions and presentations which (1) explored the current state of the art in dialog evaluation, summarizing work in a field that evolves rapidly, (2) identified the limitations of existing evaluation metrics, and (3) suggested promising research directions for future work in dialog evaluation.

This report summarizes the discussions and presentations in the workshop, with an emphasis on future research directions that funding agencies should consider going forward. The report addresses the following areas: automatic metrics in practice, assessment of evaluation metrics, human evaluation vs automatic evaluation, and the future of evaluation metrics.

1. http://dialrc.org/AED/

The current interest in a workshop on dialog evaluation metrics began in 2016, when Liu et al. (2016) empirically demonstrated that standard automatic metrics that were commonly used to evaluate natural language generation (e.g., machine translation, summarization), such as BLEU, METEOR, and ROUGE, were insufficient and indeed misleading when used to assess state-of-the-art dialog systems (Liu et al., 2016; Lowe et al., 2017). In 2018 - 2019, the interest in automatic evaluation for dialog intensified, with several papers simultaneously proposing more meaningful dialog evaluation metrics. This led to a concentrated effort to study the problem of dialog evaluation, define more meaningful automatic metrics and create datasets for the assessment of evaluation metrics. Over the past two years there has been considerable progress towards meaningful automatic evaluation metrics for dialog, with the introduction of reference-free, model-based metrics which measure multiple different dialog qualities (e.g., relevance, fluency, engagingness) and achieve moderate correlation with human judgments (Yeh et al., 2021). The newly-proposed automatic metrics are an improvement over word-overlap metrics, yet

1. they are limited in scope (they measure a limited set of dialog qualities),
2. they struggle to generalize,
3. they are not *strongly* correlated with human judgment.

As such, significant work remains to be done to improve automatic evaluation metrics for dialog. Future directions may address these three shortcomings by (1) modifying how metrics are used in practice, (2) improving upon how metrics are assessed and compared to prior work, and (3) proposing improved model architectures or training algorithms for automatic metrics. This report discusses promising directions for future research in the domain of automatic evaluation metrics for dialog.

2. Background

The evaluation of open-domain dialog is challenging. Generally, when evaluating dialog systems for specific domains or constrained settings, researchers rely on targeted evaluations, such as the evaluation of dialog state tracking (Williams et al., 2016) or task success rate (Hastie, 2012; Bordes et al., 2016). In contrast, a good evaluation metric for open-domain dialog needs go beyond these specific aspects, and must measure broader phenomena such as whether the dialog is engaging, interesting, and coherent. Two different approaches for evaluating open-domain dialog systems are used: automatic dialog evaluation metrics and human evaluation.

Though human evaluation is generally the most reliable approach, it is an expensive and time-consuming process that requires careful experimental design, as detailed in §2.1 and §3.1. Human evaluation may also suffer from inconsistency (Walker et al., 2007) and a lack of reproducibility, which is a consequence of both the annotators' subjectivity and ambiguity in the task instructions. To mitigate the difficulty and the inconsistency of human evaluation, it is imperative that the dialog research community creates meaningful automatic evaluation metrics.

Many natural language generation tasks, such as machine translation, frequently rely on automatic metrics to evaluate systems: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), or coreference resolution, a core relation extraction task using three benchmarking metrics: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and CEAF

(Luo, 2005). However, automatic evaluation of dialog is significantly more challenging, due to the inherent one-to-many nature of dialog (Zhao et al., 2017a). For a given dialog context, there are many valid responses and it is therefore impractical to select a finite set of ground-truth utterances to compare to. This makes existing NLG metrics, such as BLEU or METEOR, ineffective for assessing dialog Liu et al. (2016). To address the shortcomings of existing metrics, recent work has proposed reference-free evaluation metrics for dialog, which rely on pre-trained models to measure the appropriateness of a response rather than comparing to a ground-truth utterance. Reference-free evaluation metrics are an improvement over word-overlap metrics, however they are limited in scope, struggle to generalize and are not *strongly* correlated with human judgements. Throughout this report, we identify several issues with the current state of research on automatic evaluation metrics for dialog, and suggest future directions accordingly.

The following subsections provide more detail about both the human evaluation process and the recently proposed automatic dialog evaluation metrics.

2.1 Human Evaluation

A comprehensive analysis of the performance of dialog systems usually includes human evaluation. While automatic metrics can effectively assess certain aspects of dialog, human annotators are better able to holistically evaluate the performance of dialog systems, for example by assessing whether a generated dialog is realistic and interesting (Liu et al., 2016; Dinan et al., 2019; Adiwardana et al., 2020). Conducting human evaluation is a challenging process. Researchers must carefully consider a variety of aspects such as experimental design, number of participants, the relevant target group, task formulation, measuring scales, definition of target concepts, etc.

Trained experts usually demonstrate higher levels of agreement and better resulting sensitivity compared to annotation by crowdworkers (Gašić et al., 2013; Welleck et al., 2019; Deriu et al., 2021; Banchs, 2016). However, it is harder to recruit trained experts. Li et al. (2019a) demonstrates how changes in the phrasing of a question has an impact on the sensitivity of the study. Similarly, (Novikova et al., 2018b) show that ranking based on relative magnitudes significantly improves annotator agreement over the commonly used rating method using Likert scales. Huynh et al. (2021); Dušek et al. (2019) show that we can improve the quality of annotation by improving task description. Huynh et al. (2021) also show the impact of worker payment. They also note that tasks are more reproducible when code is made available. Recently Smith et al. (2022) has suggested that differing data collection methods have varying levels of human agreement and statistical sensitivity. Each method has a different amount of human annotation hours and labor costs. Performing human evaluation correctly, remains a challenging and open problem.

2.2 Automatic Evaluation Metrics for Dialog

Automatic evaluation metrics for dialog can be divided into two classes: referenced and reference-free metrics. Referenced metrics score a system reponse by comparing it to a reference human written utterance. These metrics are ineffective for dialog (Liu et al., 2016) due to the one-to-many nature of dialog (Zhao et al., 2017a). While this can be mitigated by using multiple reference responses (Gupta et al., 2019), it is infeasible to collect a sufficiently large dataset which thoroughly covers the space of potential responses. To this end, reference-free metrics have been proposed to circumvent the one-to-many problem. In this section, we will first introduce a few of the most pop-

ular referenced metrics, generally used as a metric for NLG tasks like machine translation. Next, we will discuss many different reference-free metrics that have been proposed to evaluate dialog.

BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) are popular rule-based metrics used to benchmark natural language generation systems. These metrics measure the word-overlap between system-generated responses and reference responses with n-gram precision and recall. These metrics are easy to use and have been widely adopted to evaluate dialog generation. However, these metrics assume the existence of ground-truth references, which is inappropriate for dialog evaluation since it is possible that completely different responses (i.e., with low word overlap) may be pragmatically relevant to the dialog context. Liu et al. (2016) have demonstrated that such metrics are ineffective for dialog evaluation due to the one-to-many nature of dialog (Zhao et al., 2017b).

ADEM (Lowe et al., 2017) and RUBER (Tao et al., 2018) are early learning-based metrics. ADEM uses a recurrent neural network to directly predict the quality of system responses. RUBER is a hybrid model consisting of both a referenced metric and an unreferenced metric. BERT-RUBER (Ghazarian et al., 2019) further improved the performance of RUBER using BERT (Devlin et al., 2019). Based on BERT-RUBER, PONE (Lan et al., 2020) uses a novel algorithm to sample negative examples during training and trains the metric on a dataset augmented by other NLG models.

Beyond RUBER, many metrics combine models measuring different dialog qualities to further improve performance. USR (Mehri and Eskenazi, 2020b) relies on language models measuring fluency, a dialog retrieval model determining the relevance of a response, and a fact-to-response selection model measuring whether the response is conditioned on knowledge information. Similarly, USL-H (Phy et al., 2020) combines models evaluating the sensibleness, likelihood, and understandability of the response. HolisticEval (Pang et al., 2020) adopts different models for evaluating several qualities of dialog: *context coherence*, *language fluency*, *response diversity*, and *logical self-consistency*. D-score (Zhang et al., 2021d) adopts a single multitask model for evaluating various dialog qualities including *context coherence*, *language fluency*, *logical self-consistency*, and *semantic appropriateness*. Deep AM-FM (Zhang et al., 2021c) measures both semantic similarity and response fluency and the PARADISE-style model of (Walker et al., 2021) uses both predicted user ratings and dialog length. DialogRPT (Gao et al., 2020) fine-tuned a GPT-2 based model (Radford et al., 2019) on Reddit human feedback data over different dimensions such as upvotes and replies. PredictiveEngage (Ghazarian et al., 2020) incorporates an utterance-level engagement classifier along with the relevance of the response to assess the overall quality of the responses. FED (Mehri and Eskenazi, 2020a) calculates the likelihood of manually-designed follow-up utterances to measure 18 fine-grained qualities of dialogs.

Researchers have designed a variety of training paradigms and model structures to further improve the model-based metrics. MAUDE (Sinha et al., 2020) is trained with Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010). DEB (Sai et al., 2020) augments training data with manually-created relevant responses and adversarial irrelevant responses and MDD-Eval (Zhang et al., 2021b) adopts a teacher model to augment dialog data across different domains to achieve better performance across domains. GRADE (Huang et al., 2020) and DynaEval (Zhang et al., 2021a) leverages a graph structure to better model the dialog. Peyrard et al. (2021) illustrate the importance of using instance-level pairing of evaluation scores. They propose mechanisms that aggregate pairwise comparisons such as estimating the probability that a given system scores better than another system. FlowScore (Li et al., 2021b) and the use of sentence embeddings in Rodríguez-Cantelar et al. (2021) model dynamic information flow in the dialog history in order to evaluate the quality of a di-

alog. FBD (Xiang et al., 2021) computes the distribution-wise difference between system generated conversations and human-written conversations to evaluate performance. AIH (Li et al., 2021a) insert questions about the facts and opinions mentioned in the dialog history in bot-bot conversations and employ human annotators or neural models to evaluate whether the responses are consistent. Adiwardana et al. (2020) proposes a new human evaluation metric called SSA and suggests perplexity on public domain social media conversations might be a good automatic metric to evaluate human qualities such as sensibleness and specificity.

In addition to metrics specifically designed for dialog evaluation, there are metrics used to evaluate general natural language generation tasks. BERTScore (Zhang et al., 2019) computes the F1 score by matching BERT embeddings between human references and system responses. BLEURT (Sellam et al., 2020) pre-trains BERT on synthetic data and fine-tunes the model to predict a human score. QuestEval (Scialom et al., 2021) accounts for factual consistency, relevance, and information selection of the generated responses. Liu et al. (2021a) uses off-the-shelf language models to create augmented responses which are used as additional references to score generated text. POSSCORE (Liu et al., 2021b) proposes an embedding-based metric taking the influence of POS tagging into account to evaluate conversational search systems.

Yeh et al. (2021) present a comprehensive study of recent automatic dialog evaluation metrics. They compare many metrics for turn-level, dialog-level, and system-level evaluation and analyze their performance on fine-grained dialog qualities. This study notes that the performance of metrics largely depends on the datasets used. It also finds that many metrics struggled with longer dialog contexts.

The fifth track of the Tenth Dialogue System Technology Challenge (Zhang et al., 2021e) was organized to address two important problems in open-domain dialog systems, each corresponding to a subtask. The first is how to design better automatic evaluation metrics to aid research and development cycles of dialog technologies. Nine teams participated in this first subtask, with similar findings to Yeh et al. (2021), for example underlining the advantage of combining multiple sub-metrics. It also demonstrated that performance on test datasets is worse than on development datasets, which highlights the need to develop robust metrics that can generalize to unseen evaluation datasets.

3. Automatic Metrics in Practice

This section explores automatic evaluation metrics from a practical perspective. First we discuss the advantages and disadvantages of human evaluation vs automatic evaluation metrics. Next, we propose the creation of annotation guidelines, which are important for standardization and facilitating collaboration within the dialog research community. Finally, we consider strategies for improving the adoption of newly proposed automatically metrics.

3.1 Tradeoffs of Human Evaluation vs Automatic Evaluations

The core advantage of human evaluation is its accurate assessment of dialog, which results from a holistic understanding of natural language. Humans are able to digest the entire dialog context in order to meaningfully evaluate a response. In contrast, automatic evaluation metrics are generally far less accurate, due to the limited ability of neural models to understand natural language dialog contexts. While some tasks in NLP have very well-defined ground-truth answers, the one-to-many problem in dialog (Zhao et al., 2017b) makes it impractical to have a *single* (or even multiple) ground-truth response. As such, during dialog evaluation, it is imperative that the evaluator (either

a human or a metric) is able to holistically understand the dialog context in order to ascertain the appropriateness of the response. As such, an advantage of human evaluation is the fact that humans are significantly better at understanding the dialog context, and therefore better able to evaluate the appropriateness of a response.

Though, human evaluation is vastly superior to automatic evaluation due to its accuracy, it does have a number of weaknesses. First, a disadvantage of human evaluation is its lack of consistency, with Walker et al. (2007) identifying a high degree of inconsistency in human judgements of automatically generated outputs, where ratings often differ significantly ($p < 0.001$) for the same utterance. Furthermore, the *same* annotator may give different scores depending on their mood, which was observed and addressed by the winning team of the Alexa Prize Socialbot Challenge 3 (Finch and Choi, 2020). This lack of consistency diminishes the reproducibility of human evaluation. In the practical sense, it may be possible to reduce inconsistencies and increase inter-annotator agreement through well-designed human evaluation tasks (Novikova et al., 2018a). Nonetheless, automatic metrics are inherently more consistent and reproducible than human evaluation. Another key shortcoming of human evaluation is its cost, which makes it impractical as a means of evaluation during training/model development (i.e., for hyperparameter tuning, model selection, etc.). Automatic evaluation is generally much cheaper and faster than human evaluation. Third, though human evaluation is generally accurate, it is difficult to carry out correctly. Designing a good evaluation task may require screening annotators and performing quality checks, especially when domain-specific knowledge is required to accurately assess the quality of responses.

This comparison of human evaluation and automatic evaluation highlights the importance of this body of research. Automatic metrics are cheaper, faster, more consistent and more reproducible. Unfortunately, they are far less accurate, largely because of the limited ability of models to holistically understand the dialog. As such, by improving upon automatic evaluation metrics to address their disadvantages, the dialog research community can build a valuable alternative to human evaluation.

User testing is a standard and well established procedure in any software development cycle. In terms of dialog systems research, human evaluation is still the most reliable and trustworthy means evaluation. As such, it is worth stressing that the objective of research into automatic metrics is not to *replace* human evaluation, but to instead *supplement* it with meaningful automatic evaluation, that is consistent, reproducible, efficient and cheap. For example, Su et al. (2016) propose an evaluator that learns online from human feedback. It queries for human feedback when under-confident in its own rating, and at the same time judges the trustworthiness of the human feedback. This results in a more robust and accurate evaluation than pure human feedback.

3.2 Guidelines for Human Annotation

An important practical aspect of dialog evaluation is consistency and reproducibility. This is particularly important for human annotation, both as a means of evaluation and to create datasets for the assessment of automatic metrics. As such, one of the first tasks to undertake before beginning human annotation is to clearly define the annotation scheme. Defining such an annotation scheme is far from straightforward, even for very narrow domains. While there are ISO standards for dialog annotation[2] (e.g., ISO 24617-2:2020 (Bunt et al., 2020) or ISO 24624:2016), these standard are difficult to follow and do not perfectly encompass the annotation challenges pertaining to evaluation.

2. Additional standards can be found at https://www.iso.org/committee/297592.html

Furthermore, hiring professional annotators is costly and does not scale, which is why researchers resort to relying on crowdworkers for annotation. As such, some benchmarks go through several iterations of annotation improvements by different groups in academia and industry – for example MultiWOZ (Budzianowski et al., 2018), MultiWOZ 2.1 (Eric et al., 2020) and MultiWOZ 2.2 (Zang et al., 2020).

Unfortunately, there is no standardized annotation scheme for dialog evaluation, particularly with respect to the dimensions to be evaluated the corresponding definitions. Often, most large-scale human annotations (i.e., to produce a human annotated dataset for assessing evaluation metrics) define a *new* annotation scheme. This is because:

- There is no unified definition or terminology for some dimensions of evaluation. Researchers may use different terms or different instructions when evaluating the same things (Finch and Choi, 2020). For example, to evaluate whether a response is fluent and natural, Li et al. (2019b) adopts the term *Fluency*, Qiu et al. (2019) uses the term *Readability*, and Zhu et al. (2019) adopts the term *Grammaticality* in the human evaluation.
- The human annotation may focus on a specific aspect (e.g., toxicity, bias, coherence, hallucination, naturalness, etc.) thereby necessitating new definitions.
- The relevant dimensions for dialog evaluation will evolve as dialog systems continue to improve. For example, early dialog systems struggled with fluency and relevance, while modern state-of-the-art systems are generally more robust in these aspects. As such, evaluation of modern systems will likely focus more on dimensions like long-term coherence or commonsense reasoning, rather than fluency.
- Each additional dimension included increases both the cost and duration of the annotation process. As such, it is likely that most human annotation experiments will drop certain dimensions.
- Introducing too many dimensions during annotation will increase cognitive load, and thereby make it more difficult for humans to evaluate a system. This may result in researchers dropping certain dimensions from evaluation, to ensure better performance on other dimensions.
- Different dimensions may be evaluated differently by different people. First of all, annotators often evaluate according to different scales. Humans may also provide different ratings depending on their cultural background, age, knowledge, experience, etc. Subjectivity in human annotation highlights the importance of thorough definitions.

Furthermore, a particular human annotation may have vastly different goals. System performance can be assessed at a turn-level, dialog-level, or at a system-level (Li et al., 2019a). Simultaneously performing human annotation for both turn-level, dialog-level and system-level evaluation will drastically increase the cost and time. Depending on how the simultaneous annotation at different levels is implemented (i.e., is turn-level annotation separate from dialog-level annotation), the quality of the dataset may be decreased due to increased cognitive load for the annotator.

Due to the lack of a standardized annotation scheme and the variety of viable goals for a particular annotation, we emphasize the importance of two takeaways for human annotation of dialog.

First, when providing a human-annotated dataset, it is important that researchers share the details of the annotation process. Human annotation for dialog evaluation is incredibly challenging,

and as such in addition to needing careful specification and definitions for each of the dimensions to be evaluated, clear guidelines and documentation are necessary in every stage of the annotation process. This includes preliminary trial annotations, inter-annotator agreements, scope of annotation, definitions and an evaluation of annotated data. The annotation scheme should indicate the annotated dimensions (ideally using the standard name for the given dimension), type of annotations, number of annotators, characteristics of the annotators, data format, etc. to facilitate the incorporation, usage and comparison of the new dataset in existing benchmarks. Gathering metadata from annotators may be a valuable means of better understanding their thought process. For instance, annotators could indicate their age or if they have experience annotating dialogs. This information may allow researchers to isolate the subjectivity in annotations, by facilitating better modelling of the annotator's expectations.

Second, it is important for future work to define standardized annotation schemes, which are well-documented and encompass many important dimensions and various viable annotation goals. To facilitate future work in the design of standardized annotation schemes, this report identifies two sets of annotation schemes that target two different levels of evaluation granularity: turn level and dialog level. The turn-level evaluation dimensions are adapted from (Finch and Choi, 2020) and presented in Table 1. The dialog-level evaluation dimensions are adapted from (Mehri and Eskenazi, 2020a) and summarized in Table 2. The respective publications include more detailed versions of the annotation schemes, particularly with respect to the dimensions. These two examples are not intended to be a proposed standardized annotation scheme for dialog evaluation, but merely an example of *effective* annotation schemes which have been used by prior works. The set of dimensions which are evaluated are particularly relevant for future work.

Dimension	**Definition**
`Grammaticality`	Responses are free of grammatical and semantic errors
`Relevance`	Responses are on-topic with the immediate dialog history
`Informativeness`	Responses produce unique and non-generic information that is specific to the dialog context
`Emotional Understanding`	Responses indicate an understanding of the user's current emotional state and provide an appropriate emotional reaction based on the current dialog context
`Engagingness`	Responses are engaging to user and fulfill the particular conversational goals implied by the user
`Consistency`	Responses do not produce information that contradicts other information known about the system
`Proactivity`	Responses actively and appropriately move the conversation along different topics
`Quality`	The overall quality of and satisfaction with the responses

Table 1: A set of turn-level evaluation dimensions adapted from (Finch and Choi, 2020)

3.3 Adoption of Metrics

Though there has been considerable progress in the development of automatic evaluation metrics for dialog in the past two years, very few of these metrics have been adopted by the general community. This raises the important question of how to encourage the research community to adopt newly developed automatic metrics. Generally, a key barrier to using new metrics is the difficulty of *relative comparison* with prior work. In order to use a new metric, a researcher would either (1) need to reproduce prior work and evaluate it with the new metric, or (2) neglect the relative comparison with prior work. Since neither of these options are desirable, researchers often opt to use more commonplace metrics since they facilitate straightforward comparisons.

Dimension	Definition
Coherence	Throughout the dialog, is the system maintaining a good conversation flow
Error Recovery	Throughout the dialog, is the system able to recover from errors that it makes
Consistency	Throughout the dialog, is the system consistent in the information it provides
Diversity	Throughout the dialog, does the system provides a diverse range of responses
Topic Depth	Throughout the dialog, does the system discuss topics in depth
Likeability	Throughout the dialog, does the system display a likeable personality
Understanding	Throughout the dialog, does the system understand the user
Informativeness	Throughout the dialog, does the system provide unique and non-generic information
Flexibility	Throughout the dialog, Is the system flexible and adaptable to the user and their interests.
Inquisitiveness	Throughout the dialog, does the system actively ask the user questions
Overall Impression	The overall quality of and satisfaction with the dialog

Table 2: A set of dialog-level evaluation dimensions adapted from (Mehri and Eskenazi, 2020a)

To deal with these issues and increase the adoption of newly proposed automatic evaluation metrics, there are two approaches which we highlight. First, competitions and leaderboards often facilitate the quick adoption of evaluation frameworks, for example DSTC9 Track 3 (Gunasekara et al., 2020) performed evaluation with the USR and FED metrics. As such, part of the onus of increasing adoption of metrics is on the organizers of challenges and the creators of corpora, as they are responsible for defining the baselines and evaluation metrics. Second, automatic metrics should be easy to use and well-documented. Repositories similar to the one[3] proposed by Yeh et al. (2021) are a vital resource to support quick adoption of new metrics. It should be relatively easy to identify the best performing evaluation metric for a particular task. This can be accomplished by building a leaderboard for automatic evaluation metrics in dialog, as outlined in §4.1.

4. Assessment of Evaluation Metrics

This section discusses several challenges and research directions pertaining to the assessment of automatic evaluation metrics. A large majority of the research on dialog evaluation aims to develop *better* automatic metrics. Currently, evaluation metrics are assessed by measuring correlation with human quality judgements. In this section, we first discuss a potential benchmark for dialog evaluation metrics. Next, we discuss the present state of datasets for dialog evaluation and highlight future directions for dialog evaluation datasets. Then, we describe fine-grained evaluation and how it is an important future direction pertaining to the assessment of metrics. Finally, we identify potential alternatives to measuring correlation with human judgments, when assessing automatic metrics.

4.1 Benchmark

In the past two years, many automatic metrics have been proposed. However, due to the fast pace of research, there was a limited amount of consistent comparisons between these different metrics. Yeh et al. (2021) performed a comprehensive assessment of many recently proposed evaluation metrics on multiple different datasets. To facilitate consistent comparisons, an important future direction is the construction of a benchmark for dialog evaluation.

Recent benchmarks and leaderboards, such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), DialoGLUE (Mehri et al., 2020), have enabled researchers to measure progress across sev-

3. https://github.com/exe1023/DialEvalMetrics

eral natural language processing tasks in a consistent manner. A benchmark like this for the assessment of evaluation metrics can similarly inspire new research, speed up advancements, and ensure consistent comparisons. Benchmarks for dialog evaluation have been proposed by SemEval (Mihalcea et al., 2019), EANCS (Wei et al., 2021), and DSTC10 track5 (Zhang et al., 2021e), though there has yet to be a permanent benchmark for evaluation. A benchmark/leaderboard may also facilitate increased adoption of new metrics, as researchers in need of an automatic evaluation metric will be able to consult the leaderboard in order to choose their metric. As such, it would be valuable to include code and documentation with each metric submitted to the leaderboard.

The organization and management of a permanent benchmark for dialog evaluation is challenging. Shortly after the launch of previous NLP benchmarks, the performance of submissions came close to the level of non-expert humans, suggesting limited headroom for further research. While we have yet to observe a similar phenomenon over the past two years of development of metrics, a benchmark for evaluation may need to be periodically expanded to encompass both new datasets, new dimensions (e.g., fluency, relevance, coherence, etc.), and new means of assessment (i.e., beyond correlation, turn-level vs dialog-level vs system-level).

A future direction strongly suggested by the attendees of this workshop is the creation of a working group that convenes regularly, creates an initial benchmark/leaderboard and coordinates assessment of evaluation metrics, as well as future developments to the leaderboard.

4.2 Datasets for Evaluation

The choice of a dataset to train or assess a metric is crucial and can make a significant impact on the conclusions we derive from any given study. Different human-annotated datasets used for assessment consist of (i) different dialog domains (e.g., topics, terminology), (ii) responses generated by different systems (e.g., older and lower-quality response generation models vs newer state-of-the-art response generation models), (iii) different dimensions (e.g., fluency, relevance, coherence), (iv) different granularities of evaluation (e.g., turn-level, dialog-level, system-level). Yeh et al. (2021) demonstrated that many recently proposed metrics did disproportionately well on datasets they evaluated on, suggesting that researchers may often inadvertently over-optimize for their target datasets. This means that many metrics are bound to perform worse on different domains, different response generation models, different dimensions and different granularities.

Since datasets play such a big role in determining the conclusions that are drawn from a particular study, future work should collect large and sufficiently diverse datasets (along the aforementioned dimensions) for assessing evaluation metrics. Such datasets should contain many domains, many response generation models, multiple different fine-grained dimensions of evaluation and multiple different granularities of evaluation. An underlying problem in dialog evaluation is the issue of generalization, i.e., that metrics should perform well on unseen data/domains. As such, an important aspect for future dialog evaluation datasets is a concertation on generalization (i.e., a dataset that will facilitate domain transfer experiments). Furthermore, the difficulty of high-quality crowdsourcing also necessitates that a good annotation scheme is developed, with clear and precise definitions for each dialog quality being evaluated.

In addition to the construction of a sufficiently large, diverse and richly annotated dataset, the participants of this workshop strongly suggest that a reference (frequently updated website, for example) be created that lists existing datasets and their properties, as well as suggested combinations of datasets for certain types of assessments. Another desirable direction is to create a group of

researchers from different sites that create the criteria mentioned above. This group would give regular updates on the website as well as at conferences attended by dialog system researchers.

4.3 Fine-grained Evaluation

Dialog quality is inherently multi-faceted (Walker et al., 1997; Mehri and Eskenazi, 2020a). There is no universal definition of *good dialog* that works in vastly different domains, applications and contexts. In certain situations, for example a task-oriented dialog system for restaurant reservations, it is important to produce factual and relevant responses and less important to produce engaging utterances. In contrast, in an open-domain chit-chat dialog system, it is more imperative to produce engaging responses. The criteria for *good dialog* changes depending on the application, the domain and even the individual users. Mehri and Eskenazi (2020b) demonstrate that different annotators value different fine-grained qualities (e.g., interesting, relevance, etc.) differently. As dialog systems and dialog evaluation metrics continue to progress, it imperative for the dialog research community to progress beyond subjective and ill-defined notions of dialog quality. To this end, we propose that future research directions should include the advancement of fine-grained evaluation.

While measuring an overall impression is often subjective and difficult to define universally, it is possible to decompose dialog quality into multiple fine-grained qualities. These fine-grained qualities, for example "relevant", "fluent", or "factual" might be defined with less subjectivity and less ambiguous definitions. We posit that a large portion of the subjectivity in dialog quality is largely a consequence of the manner in which the fine-grained qualities are combined to produce an overall impression. Decomposing dialog evaluation into these fine-grained qualities allows the same definitions and automatic metrics to be applicable in vastly different contexts. The notion of "diversity" or "interestingness" are the same in both task-oriented and open-domain dialog, however they are less important in task-oriented dialog. When using a fine-grained evaluation framework to assess a dialog system, the weighted combination of the different fine-grained qualities could be configurable based on the application and the context (e.g., certain systems do not need to be engaging).

Yeh et al. (2021) demonstrates that model-based metrics which specifically model certain dimensions (e.g., relevance, engagingness) through specific self-supervised objectives and model architectures (Mehri and Eskenazi, 2020b; Phy et al., 2020; Pang et al., 2020), achieve stronger correlations with human judgements. Generally, these metrics train sub-metrics to model specific qualities and combine these sub-metrics to produce an overall impression. Relative to the aggregated score, the fine-grained sub-metrics are often better correlated with human ratings of specific dimensions. Furthermore, combining multiple sub-metrics has been shown (Yeh et al., 2021) to better correlate with the overall impression human rating. These results empirically demonstrate the efficacy of modelling fine-grained dimensions of dialog.

The future of dialog evaluation must involve fine-grained evaluation. Datasets should be constructed that specifically annotate and measure certain dialog qualities, such as those described in §3.2. Furthermore, metrics should be independently developed for various dialog qualities. By attempting to measure an overall impression of the dialog, evaluation metrics are modelling an ill-defined phenomenon (i.e., a certain combination of fine-grained qualities that is specific to the context of the dataset used for assessment). The participants at this workshop find it to be imperative that automatic evaluation metrics specifically consider various fine-grained qualities.

4.4 Beyond Correlation with Human Judgements

Currently, automatic metrics are evaluated by measuring correlation with human judgments (Liu et al., 2016; Novikova et al., 2017). That is, the outputs of the automatic metrics are compared to human ratings with Pearson/Spearman correlation. We argue that this method of assessment may not be sufficient, as correlation with human judgements does not capture the whole picture.

A central issue with human judgments is that they are not necessarily reliable. Human studies are often under-powered (Howcroft and Rieser, 2021) and suffer from a poor conceptualisation of what they aim to measure (Howcroft et al., 2020). Clark et al. (2021) show that this disagreement can stem from insufficient annotator training; while Novikova et al. (2018b) show that annotator agreement can be improved by reducing task ambiguity, for example by using relative ranking over absolute judgement. Another source of variation could be that the disagreement is due to personal preference (Dethlefs et al., 2014). These issues with human judgements suggest the possibility that simply measuring the Spearman/Pearson correlation between automatic metrics and (averaged or otherwise aggregated) human judgements may lead to incorrect conclusions about the relative performance of different metrics. Though such a phenomenon has yet to be demonstrated empirically, it is imperative that future work studies whether measuring correlation with human judgements is a sufficient means of assessing metrics.

Recently, it was shown that comparative measurements are more robust in human evaluation. That is, for the same dialog context, the human judge would be shown different responses, and they need to rank the responses (Fomicheva et al., 2021). As such, one conceivable alternative approach for evaluating automatic metrics is to measure how well the ranking corresponds to the human rankings of candidate responses (Novikova et al., 2018b; Georgila et al., 2019b). A ranking for a set of bots would be computed from these comparisons and the metric would have to recreate it. Another approach is to acknowledge that humans will disagree on this task, e.g. due to personal preference, and consequently develop personalised evaluation metrics based on individual preference ratings, as proposed by (Dethlefs et al., 2014). These approaches, despite still comparing to human judgements, are potentially more robust and therefore may provide a more reliable signal about the quality of metric.

Furthermore, if a metric does not correlate well with human judgment, it does not necessarily imply that the metric is of low quality. Low correlation could also stem from the fact that the automatic metric and the human worker are measuring different aspects of quality. As such, a more in-depth analysis should be performed. Related to this issue is the fact that measuring correlation to human judgment does not explain why the metric differs from the human score. It does not reveal the reasoning behind the scores, which in itself is a helpful feature.

The participants of this workshop suggest that future work studies the efficacy of correlation with human judgements when assessing metrics, and explores alternatives to correlation, potentially by comparing the relative rankings of humans and metrics. Furthermore, future work should perform in-depth analysis of certain metrics to identify their shortcomings and the reasons they differ from human judgements. Another approach could be to look beyond human judgements of dialog quality, and instead focus on the impact of the interaction using extrinsic evaluation metrics (see also section 5.6). For example, the quality of a tutorial dialog system is primarily assessed through the so-called "learning gain", i.e., students take a test before and after interacting with the system and in this way it can be reliably and objectively measured how much the system has helped them learn (Core et al., 2016; Georgila et al., 2019a). Other more subjective extrinsic metrics could

be how confident or interested in a topic students feel after interacting with the system (Core et al., 2016; Georgila et al., 2019a). Likewise, for other types of dialog (e.g., recommender systems), domain-specific performance metrics can be used (e.g., how many times users follow the recommendations of the system and how often they use it). As with human judgements of dialog quality, we can calculate correlations with automated metrics or compare the relative rankings of automated metrics and extrinsic metrics.

5. Future of Automatic Evaluation Metrics

This section explores future trends in automatic metrics for dialog evaluation. We first describe the desired properties that an automated evaluation system should possess. Next, we describe each of these properties in more detail, highlighting the challenges they present. The main question posed and addressed in this section is *what are promising future directions for improving the quality of automatic evaluation metrics for dialog?*

5.1 Desired Properties

The participants in this workshop have curated a list of desired properties that researchers and practitioners should consider when designing future automatic dialog evaluation systems.

- Strong Correlation with Human Judgements (§5.2)
- Interpretability (§5.3)
- Robustness against Adversarial Attacks (§5.4)
- Generalizability (§5.5)
- Forward- and Backward-looking (§5.7)

Additional considerations besides desired metric properties include the following:

- Extrinsic Evaluation of Social Impact (§5.6)
- Human-model Collaboration (§5.8)
- Adaptation to Rapid Changes (§5.9)

5.2 Improving Correlation with Human Judgements

The goal of research on automatic evaluation metrics is to develop a suitable alternative to human evaluation, which is time-consuming, costly and difficult to reproduce. If automatic metrics are to be considered sufficient as alternatives, they should be strong predictors of human judgements. Though §4.4 argues that correlation may be a noisy means of assessment, it is still valuable to develop better correlated metrics, especially for dimensions with higher inter-annotator agreements. As such, one of the core future goals of research into automatic evaluation metrics will be to *improve correlation with human judgements*. There are several potential directions for improving correlation, some of which are enumerated below:

- **Improved models:** Model-based automatic evaluation metrics for dialog typically aim to model the relationship between the *dialog context* and the *response*. Improvements to the models can come in the form of (1) architectural improvements, (2) training algorithms, or (3) inference algorithms. Recent work has explored several promising directions of improvement including using graph-based representations of the dialog (Huang et al., 2020; Zhang et al., 2021a), combining multiple sub-metrics (Ghazarian et al., 2020; Mehri and Eskenazi, 2020b; Phy et al., 2020), better leveraging pre-trained models (Mehri and Eskenazi, 2020a; Li et al., 2021b), proposing better training objectives, and data augmentation strategies. Improving models leveraged in evaluation metrics remains a promising direction for future research.

- **Better training data:** Yeh et al. (2021) identified that many recently proposed evaluation metrics perform disproportionately well on data from the dataset they were trained on. While this is not a surprising observation, it does suggest the necessity for more diverse and representative training datasets for model-based evaluation metrics. Since most evaluation metrics are trained in a self-supervised manner, this suggests that large-scale self-supervised training may be a significant future direction for automatic evaluation for dialog. In particular, improving the generalizability of automatic evaluation metrics through better training data may significantly improve correlation with human judgement. The problem of generalizability, and potential directions to tackle it, are further addressed in §5.5.

- **Fine-grained evaluation:** As described in §4.3, fine-grained evaluation is a promising future direction for improving the correlation of automatic metrics with human judgments. There is inherent subjectivity in many human judgments, because an 'overall impression' of a response/dialog is difficult to define universally. Instead, a subjective 'overall impression' can be viewed as the weighted combination of a number of objective fine-grained qualities. While some human annotators may subjectively value *engagingness* more than *relevance*, for example, their definitions for the fine-grained qualities are more likely to be objective. As such, by developing metrics to model individual fine-grained qualities (e.g., interesting, engaging, relevant, fluent), we can achieve better correlation with human judgement. Yeh et al. (2021) demonstrates that many metrics (Mehri and Eskenazi, 2020b; Phy et al., 2020; Pang et al., 2020) which specifically model certain dimensions (e.g., relevance, engagingness) and combine multiple such sub-metrics, achieve better correlations with human judgements, in terms of both overall quality and the specific dimensions.

5.3 Interpretability

The goal of dialog evaluation is to guide researchers in developing better dialog systems. As such, interpretability is an important facet for automatic evaluation. To make automatic evaluation more interpretable, we need to develop more fine-grained and more human-centric metrics. Most existing automatic evaluation methods only give the ratings of the dialogs or of the responses. This is not sufficient to guide a developer as to how to improve systems. For example, when measuring consistency, evaluation methods usually decide whether responses contradict the context, however this binary output does not tell us the nature of the contradictions. Li et al. (2021a) propose to detect the contradictions in facts and opinions, by pointing out the specifics of the contradiction. A key future direction is to develop mechanisms of extracting more information (e.g., the reason for a particular evaluation) from automatic metrics. Freitag et al. (2021) propose a new evaluation methodology

for Machine Translation grounded in explicit error analysis, based on the Multidimensional Quality Metrics (MQM) framework, which provides insights on type and severity of error. The authors also show that automatic metrics based on word-embeddings have a surprisingly high correlation with MQM scores.

In general, there are strategies that can be employed to make black-box neural network metrics more interpretable and explainable. For example, incorporating attention mechanisms in a metric may potentially allow users of the metric to understand what part of the input is desirable or undesirable. One could take a step further and have a metric generate natural language feedback rather than simply providing numerical scores.

Furthermore, fine-grained metrics convey more information and are therefore more interpretable. As such, developing fine-grained metrics is a potential means of improving interpretability. Some existing automatic evaluation methods are designed to evaluate the overall quality (Li et al., 2021b; Huang et al., 2020; Xiang et al., 2021) or a single dimension (Ghazarian et al., 2020), but this cannot represent all of the desired dimensions of a human decision. How can we design a unified automatic evaluation metric consisting of many dimensions which are relevant to humans? Such a metric must be able to generalize to vastly different dimensions without the need for architectural/training modifications. Methods like FED (Mehri and Eskenazi, 2020a) propose to evaluate on more dimensions in a scalable manner by leveraging the implicitly learned capabilities of large-scale language models. In the interest of interpretability, future work should propose metrics which are able to easily scale to multiple dimensions, potentially through techniques like prompt tuning.

5.4 Robustness against Gaming the Metric

One under-researched area is the robustness of automatic evaluation metrics. There are different scenarios in which the automated metrics can be fooled. Sai et al. (2019) showed that trained metrics can be fooled by simple manipulations of the response. For instance, dropping punctuation or removing certain words does not decrease the scores produced by the automated metric. On the other hand, Deriu et al. (2022) showed that when an automated metric is used as a reward for reinforcement learning, the policy converges to sub optimal solutions, which are rated highly by the metric. Thus, a key future direction is to develop automatic metrics which are built to be robust against these kinds of attacks. Specifically, sub-optimal responses should not be rated highly. This future direction can be generalized to robustness against all adversarial attacks.

Besides being robust against all adversarial attacks, there are situations where the behaviors of trained metrics should not change. For instance, the metrics should be able to handle syntactic variants of the original responses, such as paraphrases or back-translated sentences. Such variants are semantically similar to the original responses. Hence, the metric scores assigned to them should be similar to those assigned to the original sentences given the same dialog context.

5.5 Metric Generalization

Existing model-based metrics have been criticized for their inability to generalize to out-of-distribution data (Yeh et al., 2021; Zhang et al., 2021b; Smith et al., 2022), i.e., they are only good at evaluating dialogs that are similar to their training data (overfit). Yet, open-domain dialog systems are becoming more versatile and exhibiting novel types of skills, such as being knowledgeable, being empathetic, and carrying out human-like daily conversations (Roller et al., 2021). To measure the quality of such conversational agents, dialog evaluation metrics are expected to conduct assessment

across different types of dialog. To improve the domain generalization of the model-based metrics, future work can consider the following two perspectives: (1) From the data perspective, we can develop more sophisticated mechanisms to generate high-quality multi-domain training data. For instance, Zhang et al. (2021b) leverages data augmentation techniques and pseudo labeling to construct a large-scale machine-annotated multi-domain dataset of around 2 million data points, called "MDD-Data". The dataset is generated based on four existing human-human dialog corpora and can be easily extended to new corpora. MDD-Data is shown to be beneficial to the multi-domain evaluation task. (2) From the algorithm perspective, we can design more sophisticated networks to explicitly target domain generalization. For example, in machine learning, meta-learning approaches, such as Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), are well-known for boosting model generalizability across domains.

Besides domain generalization, ideal evaluation metrics should exhibit good generalizability along other aspects, such as evaluation dimensions and multilingulity. Besides those suggested in table 1 and table 2, there are other dimensions that are equally important, such as addressing bias and toxicity in dialog responses (Dinan et al., 2021; Cercas Curry et al., 2021). The design of evaluation metrics should take them into consideration[4]. In addition, most of the current research works focus on evaluating English dialog. An important step should be the development of dialog evaluation datasets and automatic metrics for other languages such that the evaluation metrics can generalize to different languages.

5.6 Extrinsic Evaluation of Social Impact

Dialog evaluation should not only consider intrinsic evaluation properties, such as dialog quality, engagingness, etc., but should also consider its extrinsic social impact, such as issues related to safety, bias and stereotyping. Dinan et al. (2022) introduce a 'SafetyKit' which comprises a suite of tools and datasets to assess possible harms a dialog model could introduce. The paper addresses three safety-sensitive phenomena where ConvAI systems can cause harm: the INSTIGATOR, YEA-SAYER, and IMPOSTOR effects. In the first scenario, a dialog system generates harmful content, thereby directly instigating harm. One of the first and best known examples is the Microsoft AI chatbot "Tay", which was shut down shortly after its launch for producing offensive language (Miller et al., 2017). In the second scenario, a system may respond in a harmful manner by agreeing with (or otherwise replying unsatisfactorily to) user utterances that promote negative content, extending an approach introduced by (Lee et al., 2019). The last scenario describes situations where users receive inappropriate expert advice, e.g., medical advice (Bickmore et al., 2018). This taxonomy has already inspired further work in this area (Sun et al., 2021).

5.7 Forward-and Backward-looking Metrics

Most of the existing automatic metrics are backward-looking, which means they only look at the history of the conversation. This is reasonable for the purpose of real-time feedback. However, sometimes, we also use evaluation metrics to give delayed feedback for error analysis or model selection. Thus it is important to also look forward to the future reactions of the users to understand and assess dialog quality. Eskenazi et al. (2019) proposed to consider the *following user utterance* for both human evaluation and automatic evaluation of task-oriented dialogs. Liang et al. (2021)

4. More discussions can be found in § 5.6

proposes to use a user's feedback, such as whether the user is engaged or not, as the assessment for dialog quality. Thus, a future direction for dialog evaluation research would be: combining the traditional system-side backward looking metrics and the user-side forward looking metrics. Specifically, we can combine such metrics through simple averaging or re-weighting based on conversational tasks.

5.8 Human in the Loop

As discussed throughout this report, human evaluation is the most reliable method for assessing dialog quality but the most labor-intensive. The least labor-intensive approach would be to develop fully automated unsupervised evaluation methods, but currently these general-purpose metrics do not perform at human level, especially for complex domains.

A hybrid approach keeps humans in the loop by developing evaluation metrics using a training set labeled with human ratings (e.g., the PARADISE framework (Walker et al., 2000)). Once dialog data and human ratings are available then models can be trained to predict human ratings on unseen data (Georgila et al., 2019b, 2020). The training data is collected from dialog participants who interact with a dialog system (fully automated or Wizard of Oz-based) and labeled with ratings either from these participants or human observers who do not participate in dialog interaction but instead just read dialogs between human users and dialog systems and rate them on a variety of aspects. Typically, human observers' ratings are collected through crowdsourcing which is much more cost-effective than having actual users interact with and then rate the dialog system. Another consideration is that the training corpus will be associated with a particular version of the dialog system. A new version of the system may require collecting a new training set of dialog data and ratings. This is not ideal given that dialog system development can go through several iterations until the system is ready for deployment.

It is an open research question whether the above process can be simplified and become less expensive without sacrificing the accuracy of our prediction models. Simulated data have proven to be very popular for training dialog policies through reinforcement learning (Schatzmann et al., 2005; Georgila et al., 2006), and more recently for training deep learning-based dialog models (El Asri et al., 2016). It is much more cost-effective to generate simulated data encoding new system capabilities than deploy the new version of the system and generate a new and sufficiently large training corpus with human users, although training with human users can outperform policies trained with simulated users (Weisz et al., 2018). It is therefore important to determine whether using models trained on simulated data and human observers' ratings is a viable option. In particular, we need to investigate the resulting errors in predictions of participants' ratings when we use simulated data and observers' ratings as training data instead of participants' dialogs and participants' ratings.

Recently Georgila et al. (2020) performed an experiment addressing this question. Participants interacted with a smart-home assistant in a Wizard of Oz setting and rated their interactions with the system on a number of aspects, namely, intelligence, naturalness, personality, friendliness, their enjoyment, overall quality, and whether they would recommend the system to others. Then dialog observers rated these dialog on the same aspects. Georgila et al. (2020) also generated simulated dialogs between dialog policies and simulated users and asked observers to rate them on the same aspects. Models for predicting human ratings were trained on the simulated dialog and the observers' ratings, the Wizard of Oz dialogs and the observers' ratings, and the Wizard of Oz dialogs and the

participants' ratings. These prediction models were then applied to a held-out portion of the Wizard of Oz dialogs and tested against participants' ratings. The results of this experiment suggested that for some conversational aspects (intelligence, naturalness, overall quality) just training evaluation functions on simulated data and observers' ratings could be sufficient. This is an encouraging result which should be tested further to explore the tradeoffs involved in using simulated data versus real dialogs, and observers' ratings versus participants' ratings.

5.9 Strategy for Handling Changes in the Field

The field of dialog system assessment is evolving very rapidly. New metrics appear every few weeks and although new datasets, since they take time to collect and curate, appear with less frequency, they still are appearing at an astoundingly fast pace considering the work involved in their creation. There are several possible avenues of action for the community going forward. First, create a website that lists all of the available metrics and datasets. While a list with a link is a beginning step, it should be accompanied by an explanation of each metric and dataset, their goals and a few references to their use. The ideal webpage would have a link to some software (for example, Yeh et al. (2021)) that can be used to evaluate each metric. It should also have some metric or assessment of the breadth of use that could be expected from the given dataset.

6. Conclusions

The participants of the AED workshop have the following recommendations for future directions of research in the evaluation of dialog systems.

3.2
- Future work that performs human annotation, either for the purposes of evaluation or collecting a dataset for assessing automatic metrics, should release their annotation scheme. The annotation scheme will generally consist of (1) a set of evaluation dimensions/qualities, (2) definitions/guidelines for the annotators, and (3) documentation about the annotation procedure (metadata about the annotators, data filtering, inter-annotator agreement).
- The dialog research community should define standardized annotation schemes, which are well-documented and encompass many important dimensions and annotation goals.

3.3
- To increase adoption of metrics, the organizers of challenges/competitions and the creators of corpora should identify the best automatic metric for their dataset in order to encourage future work to rely on more meaningful automatic metrics.
- Automatic metrics should be well-documented and easy to use, potentially released as part of an all-encompassing package, similar to the HuggingFace repository[5].

4.1
- A benchmark/leaderboard should be released for automatic evaluation metrics for dialog. To make this benchmark be useful over a long period of time, it should evolve and adapt to the improvements of the dialog systems. A working group should be created that is tasked with to creating an initial benchmark, then convening regularly to coordinate assessment of metrics and extend the benchmark.

5. https://huggingface.co/

4.2
- Future work should create large and sufficiently diverse datasets for assessing evaluation metrics. Such datasets should include many dialog domains, many response generation models, multiple fine-grained qualities and multiple granularities of evaluation (i.e., turn-level, dialog-level, system-level).
- A reference should be created to list existing datasets for assessing evaluation metrics, as well as guidelines for using these datasets for assessment.

4.3
- Future research should develop metrics and collect datasets for fine-grained evaluation. The notion of the overall quality of a dialog is noisy and subjective, as such the field of automatic evaluation of dialog should move towards modelling fine-grained qualities which can be combined to form more robust and reliable evaluation metrics.

4.4
- Future work should carry out studies to assess the effectiveness of assessing metrics by measuring correlation with human judgements.
- Future work should explore alternatives to correlating metric scores and human ratings, for example comparing the relative rankings of humans and metrics.
- Future work should evaluate dialog systems and the performance of automatic metrics through extrinsic evaluations, which measure the impact of the interaction.

5.2
- Future work should improve model-based metrics through architectural improvements, training algorithms or inference algorithms. Promising directions include graph-based representations, combining multiple sub-metrics, better leveraging pre-trained models, better training objectives and data augmentation strategies.
- Future work should aim to make metrics generalizable through improved training data, i.e. by leveraging large-scale open-domain dialog data.
- Future work should develop metrics which can model different fine-grained qualities

5.3
- Future work should make metrics more interpretable, by extracting more information from automatic metrics, for example by having metrics generate explanations/justifications.
- Future work should perform in-depth analyses of certain metrics to identify their shortcomings and better explain why they differ from human judgements.

5.4
- Automatic metrics should be built to be robust against adversarial attacks, wherein system developers aim to artificially get high scores from the metric.

5.5
- Future work should explore strategies of improving the generalizability and robustness of metrics, both through multi-domain datasets and through training algorithms such as MAML.

5.6
- Evaluation should consider extrinsic social impacts of dialog technology, e.g. whether models produce output that is safe, trustworthy and bias-free.

5.7
- Future work should consider combining the traditional system-side backward looking metrics and the user-side forward looking metrics.

5.8
- Future work should explore human in the loop strategies that leverage automatic evaluation metrics.

5.9 • Future work should create additional resources (e.g. websites) for tracking new evaluation metrics and software for evaluation and datasets that have been evaluated.

7. Acknowledgement

The NSF Future Directions Workshop on Automatic Evaluation of Dialog and this report are funded by National Science Foundation grant IIS-1934222. The opinions expressed in this paper do not necessarily reflect those of the National Science Foundation.

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. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pages 563–566, 1998.

Rafael E Banchs. Expert-generated vs. crowd-sourced annotations for evaluating chatting sessions at the turn level. In *WOCHAT: Second Work-shop on Chatbots and Conversational Agent Technologies, IVA 2016*, 2016.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. Patient and consumer safety risks when using conversational assistants for medical information: An observational study of Siri, Alexa, and Google Assistant. *J Med Internet Res*, 20(9):e11510, Sep 2018. ISSN 1438-8871. doi: 10.2196/11510. URL http://www.jmir.org/2018/9/e11510/.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prevot. The iso standard for dialogue act annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558, 2020.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 7388–7403, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.587. URL https://aclanthology.org/2021.emnlp-main.587.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565.

Mark G. Core, Kallirroi Georgila, Benjamin D. Nye, Daniel Auerbach, Zhi Fei Liu, and Richard DiNinni. Learning, adaptive support, student traits, and engagement in scenario-based learning. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, Orlando, Florida, USA, 2016.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810, 2021.

Jan Deriu, Don Tuggener, Pius von Daniken, and Mark Cieliebak. Probing the robustness of trained metrics for conversational dialogue systems. *arXiv preprint arXiv:2202.13887*, 2022.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 702–711, Gothenburg, Sweden, 2014. URL http://aclweb.org/anthology/E14-1074.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*, 2019.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*, 2021.

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. SAFETYKIT: First aid for measuring safety for open-domain conversational systems. In *Proceedings of the 2022 Conference of the Chapter of the Association for Computational Linguistics (ACL)*, 2022.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002. URL https://aclanthology.org/www.mt-archive.info/HLT-2002-Doddington.pdf.

Ondrej Dušek, David Howcroft, Karin Sevegnani, and Verena Rieser. Noise and neural natural language generationrubbish in, rubbish out? In *Proceedings of EurNLP*, 2019.

Layla El Asri, Jing He, and Kaheer Suleman. A sequence-to-sequence model for user simulation in spoken dialogue systems. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1151–1155, San Francisco, California, USA, 2016.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.53.

Maxine Eskenazi, Shikib Mehri, Evgeniia Razumovskaia, and Tiancheng Zhao. Beyond turing: Intelligent agents centered on the user. *arXiv preprint arXiv:1901.06613*, 2019.

Sarah E. Finch and Jinho D. Choi. 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, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.sigdial-1.29.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. The eval4nlp shared task on explainable quality estimation: Overview and results. *arXiv preprint arXiv:2110.04392*, 2021.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation, 2021.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. Dialogue response rankingtraining with large-scale human feedback data. In *EMNLP*, 2020.

M. Gašić, C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. Young. On-line policy optimisation of bayesian spoken dialogue systems via human interaction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8367–8371, 2013. doi: 10.1109/ICASSP.2013.6639297.

Kallirroi Georgila, James Henderson, and Oliver Lemon. User simulation for spoken dialogue systems: Learning and evaluation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1065–1068, Pittsburgh, Pennsylvania, USA, 2006.

Kallirroi Georgila, Mark G. Core, Benjamin D. Nye, Shamya Karumbaiah, Daniel Auerbach, and Maya Ram. Using reinforcement learning to optimize the policies of an intelligent tutoring system for interpersonal skills training. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 737–745, Montreal, Canada, 2019a.

Kallirroi Georgila, Carla Gordon, Hyungtak Choi, Jill Boberg, Heesik Jeon, and David Traum. Toward low-cost automated evaluation metrics for Internet of Things dialogues. *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS), Lecture Notes in Electrical Engineering*, 579:161–175, 2019b.

Kallirroi Georgila, Carla Gordon, Volodymyr Yanov, and David Traum. Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 726–734, Marseille, France, 2020.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2310. URL https://aclanthology.org/W19-2310.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796, 2020.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*, 2020.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. Investigating evaluation of open-domain dialogue systems with human generated multiple references. *arXiv preprint arXiv:1907.10568*, 2019.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

Helen Hastie. Metrics and evaluation of spoken dialogue systems. In *Data-driven methods for adaptive spoken dialogue systems*, pages 131–150. Springer, 2012.

David M Howcroft and Verena Rieser. What happens if you treat ordinal ratings as interval data? human evaluations in nlp are even more under-powered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, 2021.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.inlg-1.23.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.742. URL https://aclanthology.org/2020.emnlp-main.742.

Jessica Huynh, Jeffrey Bigham, and Maxine Eskenazi. A survey of nlp-related crowdsourcing hits: what works and what does not. *arXiv preprint arXiv:2111.05241*, 2021.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1), November 2020. ISSN 1046-8188. doi: 10.1145/3423168. URL https://doi.org/10.1145/3423168.

Nayeon Lee, Andrea Madotto, and Pascale Fung. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-3655.

Margaret Li, Jason Weston, and Stephen Roller. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*, 2019a.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1002. URL https://aclanthology.org/P19-1002.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. Addressing inquiries about history: An efficient and practical framework for evaluating open-domain chatbot consistency. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1057–1067, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.91. URL https://aclanthology.org/2021.findings-acl.91.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. 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 128–138, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.11. URL https://aclanthology.org/2021.acl-long.11.

Weixin Liang, Kai-Hui Liang, and Zhou Yu. Herald: An annotation efficient method to detect user disengagement in social conversations. *arXiv preprint arXiv:2106.00162*, 2021.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1230. URL https://aclanthology.org/D16-1230.

Ruibo Liu, Jason Wei, and Soroush Vosoughi. Language model augmented relevance score. 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 6677–6690, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.521. URL https://aclanthology.org/2021.acl-long.521.

Zeyang Liu, Ke Zhou, Jiaxin Mao, and Max L Wilson. Posscore: A simple yet effective evaluation of conversational search with part of speech labelling. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1119–1129, 2021b.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1103. URL https://aclanthology.org/P17-1103.

Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1004.

Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting, July 2020a. Association for Computational Linguistics. URL https://aclanthology.org/2020.sigdial-1.28.

Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.64. URL https://aclanthology.org/2020.acl-main.64.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*, 2020.

Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang, and Soujanya Poria, editors. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/S19-1000.

K.W Miller, Marty J Wolf, and F.S. Grodzinsky. Why we should have seen that coming. *ORBIT Journal*, 1(2), Oct. 2017. doi: 10.29297/orbit.v1i2.49. URL https://www.orbit-rri.org/ojs/index.php/orbit/article/view/49.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, 2017.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2012. URL https://aclanthology.org/N18-2012.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. Rankme: Reliable human ratings for natural language generation. In *Proceedings of NAACL-HLT*, pages 72–78, 2018b.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.333. URL https://aclanthology.org/2020.acl-main.333.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. Better than average: Paired evaluation of NLP systems. 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 2301–2315, Online, August

2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.179. URL https://aclanthology.org/2021.acl-long.179.

Vitou Phy, Yang Zhao, and Akiko Aizawa. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.368. URL https://aclanthology.org/2020.coling-main.368.

Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. Are training samples correlated? learning to generate dialogue responses with multiple references. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3826–3835, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1372. URL https://aclanthology.org/P19-1372.

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

Mario Rodríguez-Cantelar, Luis Fernando D'Haro, and Fernando Matía. Automatic evaluation of non-task oriented dialog systems by using sentence embeddings projections and their dynamics. In *Conversational Dialogue Systems for the Next Decade*, pages 71–84. Springer, 2021.

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. 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, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL https://aclanthology.org/2021.eacl-main.24.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. Re-Evaluating ADEM: A Deeper Look at Scoring Dialogue Responses. In *Proceedings of the thirty-third AAAI Conference on Artificial Intelligence*, volume 33 of *AAAI'19*, pages 6220–6227, Honolulu, Hawaii, USA, 2019. URL https://aaai.org/ojs/index.php/AAAI/article/view/4581.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827, 2020. doi: 10.1162/tacl_a_00347. URL https://aclanthology.org/2020.tacl-1.52.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pages 45–54, Lisbon, Portugal, 2005.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*, 2021.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.220. URL https://aclanthology.org/2020.acl-main.220.

Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *arXiv preprint arXiv:2201.04723*, 2022.

Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2431–2441, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1230. URL https://aclanthology.org/P16-1230.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. On the safety of conversational models: Taxonomy, dataset, and benchmark, 2021.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995. URL https://aclanthology.org/M95-1005.

Marilyn Walker, Candace Kamm, and Diane Litman. Towards developing general models of usability with PARADISE. *Journal of Natural Language Engineering*, 6(3–4):363–377, 2000.

Marilyn Walker, Amanda Stent, François Mairesse, and Rashmi Prasad. Individual and domain adaptation in sentence planning for dialogue. *J. Artif. Int. Res.*, 30(1):413–456, November 2007. ISSN 1076-9757.

Marilyn Walker, Colin Harmon, James Graupera, Davan Harrison, and Steve Whittaker. Modeling performance in open-domain dialogue with paradise. *arXiv preprint arXiv:2110.11164*, 2021.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004*, 1997.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019.

Wei Wei, Bo Dai, Tuo Zhao, Lihong Li, Diyi Yang, Yun-Nung Chen, Y-Lan Boureau, Asli Celikyilmaz, Alborz Geramifard, Aman Ahuja, and Haoming Jiang, editors. *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eancs-1.0.

Gellért Weisz, Paweł Budzianowski, Pei-Hao Su, and Milica Gašić. Sample efficient deep reinforcement learning for dialogue systems with large action spaces. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2083–2097, 2018. doi: 10.1109/TASLP.2018.2851664.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Jason D Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33, 2016.

Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. Assessing dialogue systems with distribution distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2192–2198, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.193. URL https://aclanthology.org/2021.findings-acl.193.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eancs-1.3. URL https://aclanthology.org/2021.eancs-1.3.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*, 2020.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. DynaEval: Unifying turn and dialogue level evaluation. 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

5676–5689, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.441. URL https://aclanthology.org/2021.acl-long.441.

Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. MDD-Eval: Self-training on augmented data for multi-domain dialogue evaluation. *arXiv preprint arXiv:2112.07194*, 2021b.

Chen Zhang, Luis Fernando D'Haro, Rafael E Banchs, Thomas Friedrichs, and Haizhou Li. Deep AM-FM: Toolkit for automatic dialogue evaluation. In *Conversational Dialogue Systems for the Next Decade*, pages 53–69. Springer, 2021c.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021d. doi: 10.1109/TASLP.2021.3074012.

Chen Zhang, João Sadoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint arXiv:2111.02110*, 2021e.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada, July 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-1061. URL https://aclanthology.org/P17-1061.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017b.

Qingfu Zhu, Lei Cui, Wei-Nan Zhang, Furu Wei, and Ting Liu. Retrieval-enhanced adversarial training for neural response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3763–3773, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1366. URL https://aclanthology.org/P19-1366.