Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols

Sarah E. Finch
Department of Computer Science
Emory University
Atlanta, GA, USA
sfillwo@emory.edu

Jinho D. Choi
Department of Computer Science
Emory University
Atlanta, GA, USA
jinho.choi@emory.edu

Abstract

As conversational AI-based dialogue management has increasingly become a trending topic, the need for a standardized and reliable evaluation procedure grows even more pressing. The current state of affairs suggests various evaluation protocols to assess chat-oriented dialogue management systems, rendering it difficult to conduct fair comparative studies across different approaches and gain an insightful understanding of their values. To foster this research, a more robust evaluation protocol must be set in place. This paper presents a comprehensive synthesis of both automated and human evaluation methods on dialogue systems, identifying their shortcomings while accumulating evidence towards the most effective evaluation dimensions. A total of 20 papers from the last two years are surveyed to analyze three types of evaluation protocols: *automated*, *static*, and *interactive*. Finally, the evaluation dimensions used in these papers are compared against our *expert* evaluation on the system-user dialogue data collected from the Alexa Prize 2020.

1 Introduction

Most successful automated dialogue systems follow task-oriented dialogue management methodology, which defines an explicit goal that the system is seeking to fulfill through the conversation with the user (Gao et al., 2019). Recently, the research in chat-oriented dialogue management has experienced a substantial increase in popularity. Unlike task-oriented dialogues, where the success is generally measured as ability to complete the goal of the task, evaluation of chat-oriented dialogues is much less straightforward, since the conversational goals can be highly subjective (Huang et al., 2019).

The evaluation of chat-oriented dialogue systems has been typically accomplished through the use of automated metrics and human evaluation (Section 2). Automated evaluation requires no human labor once the evaluation script is written (Section 3). For automated evaluation to be a reliable measurement of the dialogue system quality, however, it needs to be shown to be a close approximation of human judgements (Section 4). Unfortunately, commonly used automated metrics correlate weakly with human judgments, indicating poor utility of such metrics (Liu et al., 2016). Human evaluation has become more commonplace in recent dialogue system works; however, it presents its own challenges. For one, it is time-consuming and expensive to obtain human judgments. More critically, there is a lack of standardized protocol for such human evaluation, which makes it challenging to compare different approaches to one another.

There have been many previous attempts at standardizing dialogue system evaluations. A major limitation has been their focus on task-oriented dialogue systems, which does not translate well to chat-oriented dialogue systems (Walker et al., 1997; Malchanau et al., 2019). Previous works which have included chat-oriented evaluations have lacked comprehensive coverage over the many varieties of such evaluation procedures that are currently in use. Instead, the emphasis has rested primarily on automated metrics at the expense of detailed analysis of human evaluation (Deriu et al., 2019). At this stage in conversational AI, it is probable that automated and human metrics reveal different aspects of dialogue systems (Hashimoto et al., 2019). It would be remiss to focus on a single evaluation category when assessing the state of the field. For this reason, our work aims to fill in the gaps of previous dialogue system evaluation surveys by identifying and comparing human evaluation protocols for chat-oriented dialogue systems.

To this end, we present a comparative analysis of the evaluations used for chat-oriented dialogue systems over the past several years. Since the field of conversational AI has experienced a rapid growth

, pages 243 - 252

in these years, it presents a unique opportunity to observe and assess which evaluation metrics have been most widely adopted by the larger community in this period of expeditious development. We provide a detailed survey of both automated and human evaluations in order to present the most accurate depiction of the current evaluation protocols. However, our in-depth analysis is limited to that of the human evaluations due to the abundance of previous work in automated metric analysis. As such, we defer to such work as Liu et al. (2016), Ghandeharioun et al. (2019), and Ghazarian et al. (2019) for more detail on automated metrics.

As a part of our analysis, we also present a case study of real human-machine dialogues which explores the significance of different human evaluation metrics in terms of overall user satisfaction through an expert analysis. As a result of our work, the most commonly used evaluation metrics in contemporary literature - both automated and human - are revealed in detail and our findings towards the prevalence, impact, and applicability of human evaluation metrics are illustrated.

2 Evaluation Protocols

For a holistic understanding of current evaluation protocols on dialogue systems, we have carefully selected 20 relevant papers since 2018, primarily from top-tier venues, and synthesized their methods. These papers focus on open domain (or non-task-oriented) dialogue, and employ a variety of approaches including:[1]

- Incorporation of knowledge bases [2, 4, 7, 18, 20]
- Integration of personality [8, 12]
- Handling of emotion-driven responses [10]
- Purely depending on neural-based sequence-to-sequence models [19]

Based on these papers, three main categories are found as evaluation protocols for open-domain dialogue systems: *automated*, *static*, and *interactive*. Automated evaluation is performed systematically by a batch script such that no human effort is required once the script is written (Section 2.1). Static evaluation is done by human where the evaluator assesses a dialogue whose last utterance is generated by the dialogue system (Section 2.2). Interactive evaluation is also done by human, although the evaluator assesses the quality of the dialogue after directly interacting with the dialogue system (Section 2.3).

Table 1 shows the distributions of the three evaluation protocols. Most recent approaches adopt both automated and human evaluations, with only 2 papers not including any form of human evaluation. The most common protocol for human evaluation is static evaluation, with very few papers conducting interactive assessments of dialogue systems. No work has adopted all three types of evaluation protocols.

Method	References	#
AUT	[1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12 13, 14, 15, 16, 17, 20]	17
STA	[1, 3, 4, 7, 9, 10, 11, 12, 13, 14 15, 16, 17, 18, 19, 20]	16
INT	[8, 19]	2
AUT & STA	[1, 3, 4, 7, 9, 10, 11, 12, 13, 14 15, 16, 17, 20]	14
AUT & INT	[ ]	0
STA & INT	[19]	1

Table 1: Distributions of the three evaluation protocols. #: number of papers using the corresponding protocol, AUT/STA/INT: automated/static/interactive evaluation. &: approaches using both protocols.

2.1 Automated Evaluation

Automated evaluation provides an objective quantitative measurement of the dialogue systems by operationalizing various dimensions of dialogue into mathematical formulations. Depending on the specific objectives behind different systems, a few studies define novel automated metrics to capture the benefit of their proposed approaches. Automated evaluation provides the most straightforward and undemanding methods by which to evaluate dialogue systems; however, they are generally viewed as poor indicators of true dialogue quality, following results from Liu et al. (2016).

2.2 Static Evaluation

Static evaluation is an offline procedure where the evaluators never directly interact with the dialogue systems under review; instead, they are provided with dialogue excerpts. These excerpts are gen-

[1]Throughout the paper, the following are used to refer to the related work: 1: Li and Sun (2018) 2: Liu et al. (2018) 3: Luo et al. (2018) 4: Moghe et al. (2018) 5: Parthasarathi and Pineau (2018) 6: Xu et al. (2018) 7: Young et al. (2018) 8: Zhang et al. (2018) 9: Du and Black (2019) 10: Li et al. (2019) 11: Lin et al. (2019) 12: Madotto et al. (2019) 13: Qiu et al. (2019) 14: Tian et al. (2019) 15: Wu et al. (2019) 16: Zhang et al. (2019) 17: Zhou et al. (2019) 18: Zhu et al. (2019) 19: Adiwardana et al. (2020) 20: Wang et al. (2020).

erated by first randomly sampling dialogues from a corpus consisting of human-to-human conversations, then having the systems produce responses to the sampled dialogues. The sampled dialogues together with the system responses are provided to human evaluators to assess. Because only the last utterance in these excerpts are generated by the dialogue systems, it is difficult to evaluate sequential aspects about dialogue management through static evaluation (e.g., coherence among responses generated by the same system).

2.3 Interactive Evaluation

Unlike static evaluation, interactive evaluation has the same person play the role of both the user (one who interacts with the system) and the evaluator. In this setup, the evaluator has a conversation with the dialogue system and makes the assessment at the end of the conversation. Even though this procedure is more demanding in terms of time and human effort than static evaluation, it allows the evaluator to gain a better sense of the capability of the dialogue system through explicit interaction.

3 Analysis of Automated Evaluation

Table 2 shows the 11 metrics used for automated evaluation in our survey:

- `BLEU`: a subset of BLEU-1 through BLEU-4 (Papineni et al., 2002)
- `C`: sum of entailment scores between response and persona description (Madotto et al., 2019)
- `Coherence`: average word embedding similarity between dialogue context and generated response (Xu et al., 2018)
- `Distinct`: a subset of Distinct-1, Distinct-2, and Distinct-sentence (Li et al., 2016)
- `Embedding`: a subset of average, extrema, and greedy embedding similarity (Liu et al., 2016)
- `Entity A/R`: Accuracy and recall for including the correct entities in the response (Liu et al., 2018)
- `Entity Score`: average number of entities per response (Young et al., 2018)
- `Entropy`: average character-level entropy over all responses (Mou et al., 2016)
- `Inertia`: inertia on the clusters of embeddings of responses (Du and Black, 2019)
- `Perplexity`: inverse likelihood of predicting the responses of the test set (Chen et al., 1998)
- `ROUGE`: a subset of ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004)

The automated metrics in Table 2 fall into the following five categories:

Ground Truth Response Similarity Most commonly used automated metrics focus on assessing how well system responses match the ground truth human responses, using word overlap (`BLEU`, `ROUGE`) or embedding similarity.

Context Coherence Embedding similarities between dialogue contexts and system responses have been used to quantitatively assess the relevance between the system responses and the preceding dialogue history (`Coherence`, `Embedding`).

Response Diversity Other widespread metrics assess the diversity of the system responses in order to determine the amount of repetition and generic content in the system responses (`Distinct`, `Entropy`, `Inertia`, `Entity Score`).

Language Model Fitness Generative models are usually evaluated in terms of how well they learn to model the language of the dialogues in their training corpus (`Perplexity`).

Application-Specific The other observed metrics can be considered application-specific since `Entity A/R` is used to measure the ability of the system to produce the correct entities in its responses and `C` is specifically created as a measure of the consistency between the dialogue responses and their respective persona descriptions.

4 Analysis of Human Evaluation

While automated evaluation measures dimensions of dialogue objectively, human evaluation captures the subjective assessment from the user's point of view. Regardless of the exact method chosen, all human evaluations involve gathering external annotators who answer questions regarding the dialogues resulting from a dialogue system.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	#
BLEU			✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	14
C												✓									1
Coherence						✓															1
Distinct	✓		✓			✓			✓				✓	✓	✓	✓	✓				9
Embedding	✓								✓				✓	✓			✓				5
Entity A/R		✓																			1
Entity Score							✓													✓	2
Entropy														✓							1
Inertia									✓												1
Perplexity					✓		✓			✓		✓			✓	✓				✓	7
ROUGE				✓					✓												2

Table 2: Metrics of the automated evaluation used by recent papers on open-domain dialogue systems. The top row shows the reference numbers to the 20 surveyed papers. #: number of papers using the corresponding metrics.

	1	2	3	4	7	8	10	11	12	13	14	15	17	18	19	20	#
Appropriateness					✓									✓			2
Coherence			✓									✓					2
Consistency	✓					✓			✓								3
Context Coherence							✓										1
Correctness		✓														✓	2
Diversity										✓							1
Emotion	✓																1
Empathy								✓									1
Engagingness						✓											1
Fluency		✓	✓	✓		✓	✓	✓	✓			✓				✓	9
Grammaticality														✓			1
Humanness				✓													1
Informativeness					✓						✓	✓		✓			4
Knowledge Rel.		✓					✓									✓	3
Logic	✓																1
Proactivity												✓					1
Quality											✓		✓				2
Readability										✓							1
Relevance				✓				✓		✓							3
Sensibleness															✓		1
Specificity				✓											✓		2

Table 3: Dimensions of the human evaluation used by recent dialogue system papers. The top row shows the reference numbers to the 20 survey papers. [5, 6] do not perform any human evaluation; [9, 16] perform human evaluation without reference to dimensions. #: number of papers adopting the corresponding dimensions.

4.1 Dimensions of Human Evaluation

There is high variability in the dimensions of dialogue that previous studies have used for assessing dialogue systems in both static and interactive evaluations. Table 3 provides a detailed overview of the dimensions used by each of the surveyed papers when evaluating their work. There are a total of 21 uniquely-worded dimensions found; 11 of them appear in only a single paper. The resulting matrix provides clear evidence of the inconsistencies in human evaluation methods, as its sparsity is indicative of low overlap among those methods. The long tail distribution of the evaluation metrics makes it difficult for cross-work comparisons without a substantial study to align the disparate evaluation of one work with another.

Although the evaluation dimensions appear to be distinct on the surface, several of them appear to be similar in meaning. To analyze the level of overlap among the seemingly distinct evaluation dimensions, we compile the definitions and instructions shared by each of the papers regarding their evaluation dimensions and rating scales. Based on manual analysis, we are able to group dimensions together that are indeed evaluating the same aspect of dialogue as one another, even though the authors mention them by different names. Table 4 provides the dimension groupings that are identified on the basis of their respective definitions.

Definitions in Table 4a aim to address the grammaticality of system responses, including words like *grammar*, *understandable*, and *accurate*. As

Fluency	Whether the response from the listener is understandable (Lin et al., 2019)
	Whether the response is fluent and natural (Li et al., 2019)
	Whether each sentence has correct grammar (Luo et al., 2018)
	Fluency measures if the produced response itself is fluent (Wu et al., 2019):
Consistency	Whether the reply is fluent and grammatical (Li and Sun, 2018)
Readability	Whether the utterance is grammatically formed (Qiu et al., 2019)
Grammaticality	Whether the response is fluent and grammatical (Zhu et al., 2019)

(a) Grammatical Capability.

Relevance	Whether the responses of the listener seem appropriate to the conversation (Lin et al., 2019)
	Whether the response is appropriate/relevant in the current context language (Moghe et al., 2018)
	Whether the reply is relevant to the query (Qiu et al., 2019)
Appropriateness	Whether the response is appropriate in grammar, topic, and logic (Young et al., 2018)
Coherence	Whether the generated response is relevant to the input (Luo et al., 2018)
	Whether the whole dialogue is fluent (does not contain irrelevant or illogical responses) (Wu et al., 2019)
Context Coherence	Whether the response is coherent with the context and guides the following utterances (Li et al., 2019)
Logic	Whether the post and the reply are logically matched (Li and Sun, 2018)
Sensibleness	Whether the response makes sense given the context (Adiwardana et al., 2020)

(b) Turn Coherence.

Informativeness	Whether the response provides new information and knowledge in addition to the post (Young et al., 2018)
	Whether the response has unique words and multi-topic clauses (Tian et al., 2019)
	Whether the response has meaningful information relevant to its message (Zhu et al., 2019)
	Whether the model makes full use of knowledge in the response (Wu et al., 2019)
Specificity	Whether the model produced movie-specific responses or generic responses (Moghe et al., 2018)
	Whether the response is specific to the context (Adiwardana et al., 2020)
Diversity	Whether the reply narrates with diverse words (Qiu et al., 2019)

(c) Response Informativeness.

Table 4: Proposed reductions of dialogue evaluation dimensions into non-overlapping components

a result, the four dimensions recorded in this table can be viewed as lexical variations of the same underlying Grammaticality dimension. Similarly, definitions in Table 4b highlight keywords like *appropriate*, *relevant*, and *on-topic*, thus providing evidence that each of those dimensions are instances of the Relevance dimension. Finally, Table 4c has a high occurrence of information and diversity-focused definitions, and we can reduce the dimensions shown there to the single Informativeness dimension.

Other than these highly overlapping dimensions, Quality (Tian et al., 2019; Zhou et al., 2019) and Humanness (Moghe et al., 2018) can both be considered as the single Quality dimension, since they are used to elicit an overall quality assessment of the dialogue system responses. Similarly, Emotion (Li and Sun, 2018) and Empathy (Lin et al., 2019) can be reduced into the Emotional Understanding dimension that captures both the comprehension and production of emotional responses. The remaining two dialogue dimensions assess a unique quality of dialogue and are useful as independent dialogue dimensions:

- Engagingness: whether the response includes interesting content (Zhang et al., 2018)
- Proactivity: whether the response introduces new topics without breaking coherence (Wu et al., 2019)

Finally, two evaluation dimensions are specifically used for a subset of dialogue systems that incorporate knowledge:

- Correctness: was the response accurate based on the real-world knowledge (Liu et al., 2018; Wang et al., 2020)
- Knowledge Relevance: was the knowledge shared in the response appropriate to the context (Liu et al., 2018; Wang et al., 2020)

Knowledge Relevance is very similar to the previously discussed Relevance dimension, although it is specifically targeting an assessment of the appropriateness of the knowledge being used. Even more niche, the Correctness dimension is unique to knowledge-focused systems that seek to present only true factual information to the user; thus, such a dimension may not be useful in other contexts. Due to their targeted nature, these two dimensions may fall outside of the scope of a general, comprehensive, unified evaluation of dialogue systems, and instead be used for a targeted subgroup.

Dimension	Definition
Grammaticality	Responses are free of grammatical and semantic errors
Relevance	Responses are on-topic with the immediate dialogue history
Informativeness	Responses produce unique and non-generic information that is specific to the dialogue context
Emotional Understanding	Responses indicate an understanding of the user's current emotional state and provide an appropriate emotional reaction based on the current dialogue 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 dialogue

Table 5: The final set of our proposed dialogue dimensions for human evaluation.

In total, after merging similar dimensions and discarding non-generalizable dimensions, a total of eight dimensions have been identified that share little to no definitional overlap and are reasonably applicable to all dialogue systems. Table 5 shows the finalized set of dialogue evaluation dimensions.

4.2 Diversities in Evaluation Metrics

Aside from the discrepancies in dialogue dimensions used for evaluation among different works, the actual procedure of evaluating these dialogue dimensions varies even further, particularly for static evaluations. A majority of work instructs human annotators to rate the dialogue system responses on a set of dialogue dimensions using numeric scales, where the scales being used are often different even between works that employ the same dialogue dimensions. For instance, one of the most commonly used dimension is the Fluency of the dialogue, with 9 out of the 16 papers in Table 3 have adopted this as an evaluation dimension. Between those 9 studies, Fluency ratings include scales of:

- 0∼2: Wu et al. (2019); Li et al. (2019)
- 0∼3: Wang et al. (2020); Liu et al. (2018)
- 1∼5: Moghe et al. (2018); Zhang et al. (2018); Lin et al. (2019); Madotto et al. (2019)
- 1∼10: Luo et al. (2018)

Furthermore, some studies use a preference metric for static evaluation in addition to - or even instead of - the numerical ratings (Lin et al., 2019; Young et al., 2018; Du and Black, 2019; Zhang et al., 2019). In this case, human annotators are asked to select the most compelling response among many generated by multiple dialogue systems or even humans. Thus, preference metrics provide estimated ranking scores among different systems by measuring the percentage of times each system is preferred over the others.

Unlike the diversity in static evaluation, for the two papers, Zhang et al. (2018) and Adiwardana et al. (2020), employing interactive evaluation, only numerical ratings on specific dialogue dimensions are used as evaluation methods; other methods such as preference metrics are not used in either case.

4.3 Static vs Interactive Evaluations

Establishing the necessary assessment metrics is only one consideration to achieve an accurate dialogue evaluation. The other major consideration is the procedure underlying the evaluation. This section discusses the two human evaluation protocols, static and interactive evaluations, that have previously been used by many dialogue systems. Although both evaluation protocols overcome the deficiencies brought forth by automated evaluation through human judgment, interactive evaluation is hypothesized to be a more reliable assessment strategy than static one. What static evaluation offers above interactive evaluation is a lower cost in terms of time and labor. By removing the human annotator from the task of interacting with the dialogue system, and instead having them review a dialogue excerpt, the amount of work required is reduced.

However, this is simultaneously a point in favor of static evaluation, but also a factor as to why it is less reliable. As Ghandeharioun et al. (2019) suggest, chat-oriented dialogues have a less defined conversational goal which can best be summarized as being able to hold a "natural social interaction with humans". The success - or failure - at this can only be evaluated by the targeted recipient of the conversation; namely, the user that the system is interacting with. External annotators, at best, can estimate the user's satisfaction with the conversation based on their own projected opinions, which is not necessarily the most accurate assessment.

In addition, static evaluation is commonly conducted by producing a single system response in

OQ	GR	RE	IN	EU	EN	CO	PR
1	5.00 (±0.00)	1.94 (±0.98)	2.86 (±1.29)	1.00 (±0.00)	2.33 (±0.89)	4.94 (±0.23)	1.64 (±0.87)
2	4.70 (±0.47)	2.85 (±0.88)	3.25 (±1.25)	1.15 (±0.37)	3.15 (±0.75)	4.90 (±0.31)	2.15 (±0.59)
3	4.62 (±0.51)	3.46 (±0.52)	2.92 (±0.86)	1.08 (±0.28)	2.92 (±0.49)	4.77 (±0.44)	2.38 (±0.65)
4	4.71 (±0.46)	3.89 (±0.42)	4.25 (±0.70)	1.11 (±0.31)	3.86 (±0.36)	4.82 (±0.39)	2.93 (±0.54)
5	4.33 (±0.58)	4.33 (±0.58)	3.67 (±0.58)	1.33 (±0.58)	4.00 (±0.00)	5.00 (±0.00)	3.00 (±0.00)

(a) The OQ column shows the overall quality ratings from our expert and the other columns show the average ratings from the expert on the corresponding dialogue dimensions.

OQ	GR	RE	IN	EU	EN	CO	PR
1	4.85 (±0.37)	2.20 (±1.20)	2.95 (±1.28)	1.00 (±0.00)	2.60 (±1.05)	4.85 (±0.37)	1.95 (±0.94)
2	4.80 (±0.41)	3.05 (±1.10)	3.95 (±1.19)	1.25 (±0.44)	3.30 (±0.92)	5.00 (±0.00)	2.10 (±0.79)
3	4.85 (±0.37)	2.75 (±1.07)	2.50 (±0.95)	1.00 (±0.00)	2.60 (±0.75)	4.90 (±0.31)	2.05 (±0.89)
4	4.65 (±0.49)	3.40 (±0.82)	3.30 (±0.92)	1.10 (±0.31)	3.25 (±0.79)	4.85 (±0.37)	2.25 (±0.72)
5	4.80 (±0.41)	3.30 (±1.13)	4.10 (±0.97)	1.05 (±0.22)	3.50 (±0.76)	4.80 (±0.41)	2.85 (±0.75)

(b) The OQ column shows the overall quality ratings from the Alexa Prize and the other columns show the average ratings from the expert on the corresponding dialogue dimensions.

Table 6: The average ratings by our expert on each of the dialogue dimensions in Table 5 with respect to the overall ratings from the expert and the Alexa Prize. OQ: Quality, GR: Grammaticality, RE: Relevance, IN: Informativeness, EU: Emotional Understanding, EN: Engagingness, CO: Consistency, PR: Proactivity.

a fixed dialogue context. This fails to reveal certain system deficiencies, such as repetitiveness, inconsistency, and lack of long-term memory of the information shared in the conversation. It also prevents an assessment of the system's error-handling or misunderstanding recovery capabilities from being encountered. All of these aspects are necessary to truly assess the quality of dialogues that a given dialogue system can produce. Without this information, only a biased perspective can be achieved, and the evaluation will not reflect the true capability of the system if it were to be used in practice.

5 Case Study: Alexa Prize 2020

This section presents a case study of the significance of the proposed dialogue dimensions in Table 5 using real human-machine dialogues. For this analysis, 100 rated conversations were taken from the Alexa Prize Socialbot Grand Challenge 3[2], which is a university competition to create innovative open-domain chatbots (Ram et al., 2018). During the competition, conversations are rated in terms of `Overall Quality` on a scale of `1 (worst)` to `5 (best)` under the interactive evaluation protocol. For this case study, we sampled conversations with an equal distribution between all ratings, where every conversation has at least three turns to ensure sufficient content.

Because only the `Overall Quality` dimension is provided from the interactive evaluation, we also conducted an expert analysis on the same conversations in order to explore the implications of the other previously identified dialogue dimensions. To this end, one of the authors - who has over three years of experience in dialogue system research - manually rated the conversations on each of the dialogue dimensions in Table 5.

It is worth mentioning that the following findings are taken as only a preliminary analysis, strongly considering the low agreement between the expert and interactive evaluations on OQ, which will be discussed shortly (Section 5.2). This disparity between the expert and human user evaluations renders it difficult to convey a convincing conclusion regarding the significance of the evaluation dimensions. However, we hope this work begins the momentum to investigate the importance of such evaluation dimensions in overall human perception of dialogue quality.

5.1 Quality vs. Other Dialogue Dimensions

Table 6 shows the average rating and its standard deviation on each of the 7 dialogue dimensions (GR, RE, IN, EU, EN, CO, PR) across the overall quality ratings (OQ). All ratings on those 7 dimensions are assessed by our expert. OQ ratings are provided by the expert for Tables 6a and the human users from the Alexa Prize for Table 6b.

Relevance & Proactivity The clearest positive relationship to OQ is observed from RE and PR, especially from the expert evaluation although it can be seen in the interactive evaluation as well. This suggests that these dimensions are pertinent to the human perception of dialogue quality, and that this relationship is even more apparent when evalu-

[2] https://developer.amazon.com/alexaprize

ators are given the opportunity to review previous dialogue turns when determining OQ.

Informativeness & Engagingness The relation ship between IN and EN to OQ is not as obvious as the previous two dimensions, RE and PR, although an indication of a positive relationship is observed.

Grammaticality Due to the manual curation of responses in our Alexa Prize chatbot, we have tight control over the grammaticality of our responses; thus, the overall variance in GR is low. Interestingly, we do notice that there is a slight inverse relationship between GR and OQ. Although this may seem counter-intuitive, the likely explanation is that conversations with higher OQ tend to be longer so that they comprise a greater number of topics and, as more topics are introduced, the chance for an (accidentally) ungrammatical response to be revealed is higher. Nonetheless, it appears that ungrammaticality is not a strict deterrent on OQ.

Emotional Understanding & Consistency The effect of EU and CO on OQ is inconclusive from the presented analysis. This is attributed to the low variation in these dimensions of our chatbot, as we can enforce the consistency of responses and do not aim to tackle emotional understanding.

5.2 Expert vs. Interactive Evaluations

The inter-annotator agreement between the OQ ratings of the expert and the users from the Alexa Prize is provided in Table 7. The agreement is measured for both fine-grained ratings that consider all scales (1 - 5) and coarse-grained ratings that consider only two scales (low: 1 - 2, high: 3 - 5). Although the inter-annotator agreement is higher for the coarse-grained ratings, it is apparent that the agreement scores are dramatically low for both.

Rating Type	Agreement
Fine-grained	0.13
Coarse-grained	0.22

Table 7: Cohen's Kappa scores on the overall quality ratings between the expert and interactive evaluation.

Table 8 shows that the expert evaluation tends to be more punishing overall, with a much fewer number of conversations receiving a 5.0 rating. Indeed, 56% of the conversations from the expert evaluation would be categorized as a low rating, whereas the interactive evaluation has only 40%. Even so, the low agreement indicates that the quality assessments across the two evaluation protocols are highly variable across the same conversations.

OQ	1	2	3	4	5	$\sum$
Interactive	20	20	20	20	20	100
Expert	36	20	13	28	3	100

Table 8: Comparison of the rating distribution between expert and interactive evaluation

This provides preliminary support for the hypothesis in Section 4 that external evaluators are unable to accurately infer the same impression of a conversation as that of the user who is actually participating in the conversation. Although there are potential methods which aim to mitigate this effect - such as agglomerate ratings across more than one external annotator - the underlying cause of such variance may be attributed to the poor suitability of external evaluations for dialogue system evaluation as a whole, but further work is required.

6 Conclusion and Future Work

In this paper, we provide an extensive background and the current states on the three types of dialogue system evaluation protocols, automated, static, and interactive. Our analysis shows that static evaluation is the dominating human evaluation used in the most recent dialogue system works, although it has several concerning limitations, some of which are exemplified through our case study. We propose a set of eight dialogue dimensions that encapsulate the evaluations of previous studies without redundancy. As a result of our case study, we find preliminary evidence that the dimensions of relevance, proactivity, informativeness, and engagingness are likely to be contributing factors to the overall perception of dialogue quality.

Our future work will build upon these findings to develop a thorough understanding of the necessary dialogue dimensions for comprehensive interactive evaluation of dialogue systems. Through an analysis based on large-scale user studies, we look to propose an evaluation protocol that captures the human judgement of dialogue quality through precise formulation of evaluation dimensions, in order to enable targeted dialogue system advancements.

Acknowledgments

We gratefully acknowledge the support of the Alexa Prize Socialbot Grand Challenge 3. Any contents in this material are those of the authors and do not necessarily reflect the views of the Alexa Prize.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*.

Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. 1998. Evaluation metrics for language models. In *DARPA Broadcast News Transcription and Understanding Workshop (BNTUW)*.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on evaluation methods for dialogue systems. *arXiv preprint arXiv:1905.04071*.

Wenchao Du and Alan W Black. 2019. Boosting Dialog Response Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 38–43, Florence, Italy. Association for Computational Linguistics.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems. In *Advances in Neural Information Processing Systems 32*, pages 13658–13669. Curran Associates, Inc.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. 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. Association for Computational Linguistics.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. 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 1689–1701.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2019. Challenges in building intelligent open-domain dialog systems. *arXiv preprint arXiv:1905.05709*.

Jingyuan Li and Xiao Sun. 2018. A Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 678–683, Brussels, Belgium. Association for Computational Linguistics.

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

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 56–60, Barcelona, Spain. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of Empathetic Listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge Diffusion for Neural Dialogue Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Liangchen Luo, Jingjing Xu, Junyang Lin, Qi Zeng, and Xu Sun. 2018. An Auto-Encoder Matching Model for Learning Utterance-Level Semantic Dependency in Dialogue Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 702–707, Brussels, Belgium. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing Dialogue Agents via Meta-Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Andrei Malchanau, Volha Petukhova, and Harry Bunt. 2019. Multimodal Dialogue System Evaluation: A Case Study Applying Usability Standards. In *9th International Workshop on Spoken Dialogue System Technology*, volume 579, pages 145–159. Springer Singapore, Singapore.

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

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending Neural Generative Conversational Model using External Knowledge Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. 2019. 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. Association for Computational Linguistics.

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

Zhiliang Tian, Wei Bi, Xiaopeng Li, and Nevin L. Zhang. 2019. Learning to Abstract for Memory-augmented Conversational Response Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3816–3825, Florence, Italy. Association for Computational Linguistics.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving Knowledge-aware Dialogue Generation via Knowledge Base Question Answering. *arXiv preprint arXiv:1912.07491*.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive Human-Machine Conversation with Explicit Conversation Goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better Conversations by Modeling, Filtering, and Optimizing for Coherence and Diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, Brussels, Belgium. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4970–4977.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. ReCoSa: Detecting the Relevant Contexts with Self-Attention for Multi-turn Dialogue Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730, Florence, Italy. Association for Computational Linguistics.

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

Kun Zhou, Kai Zhang, Yu Wu, Shujie Liu, and Jingsong Yu. 2019. Unsupervised Context Rewriting for Open Domain Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1834–1844, Hong Kong, China. Association for Computational Linguistics.

Qingfu Zhu, Lei Cui, Wei-Nan Zhang, Furu Wei, and Ting Liu. 2019. 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. Association for Computational Linguistics.