Toward More Accurate and Generalizable Evaluation Metrics for Task-Oriented Dialogs

Abishek Komma, Nagesh Panyam Chandrasekarasastry, Timothy Leffel
Anuj Goyal, Angeliki Metallinou, Spyros Matsoukas, Aram Galstyan

Amazon Alexa AI

{kommaak,nagecha,leffelt,anujgoya,ametalli,matsouka,argalsty}@amazon.com

Abstract

Measurement of interaction quality is a critical task for the improvement of spoken dialog systems. Existing approaches to dialog quality estimation either focus on evaluating the quality of individual turns, or collect dialog-level quality measurements from end users immediately following an interaction. In contrast to these approaches, we introduce a new dialog-level annotation workflow called Dialog Quality Annotation (DQA). DQA expert annotators evaluate the quality of dialogs as a whole, and also label dialogs for attributes such as goal completion and user sentiment. In this contribution, we show that: (*i*) while dialog quality cannot be completely decomposed into dialog-level attributes, there is a strong relationship between some objective dialog attributes and judgments of dialog quality; (*ii*) for the task of dialog-level quality estimation, a supervised model trained on dialog-level annotations outperforms methods based purely on aggregating turn-level features; and (*iii*) the proposed evaluation model shows better domain generalization ability compared to the baselines. On the basis of these results, we argue that having high-quality human-annotated data is an important component of evaluating interaction quality for large industrial-scale voice assistant platforms.

1 Introduction

Automated measurement of interaction quality is a critical task for the development and improvement of large-scale voice-based AI assistants. There has been a substantial amount of recent work on automated dialog evaluation both for open domain (Ji et al. 2022; Ghazarian et al. 2021; Ghazarian et al. 2022a) and task-oriented (Bordes et al. 2017; Lubis et al. 2022) dialog systems (for recent surveys, see Deriu et al. 2021; Yeh et al. 2021). For task-oriented dialog (TOD) systems such as conversational AI assistants, existing research has largely focused on evaluating the quality of individual turns (Ultes et al. 2014; Schmitt and Ultes 2015; Gupta et al. 2021; a.o.). Estimating interaction quality at the multi-turn dialog level is a relatively less studied topic (though see Walker et al. 1997; Walker et al. 2000; Bodigutla et al. 2020; Deng et al. 2022).

Bodigutla et al. (2019a); Bodigutla et al. (2020) showed that user-supplied (binarized) dialog-level satisfaction ratings can be predicted with 70-80% accuracy on a multi-domain dataset, if turn-level quality predictions are included as input features in a supervised model. However, in real-world scenarios, user-satisfaction ratings can be logistically difficult—and frustrating for users—to gather at a sufficiently large scale. Such ratings are also influenced by factors irrelevant to interaction quality itself, e.g. novice and expert users might rate the same dialog very differently; see Bodigutla et al., 2019a for discussion.

Turn	User utterance	System response
1	what is the score of the giants game	the new york giants are trailing the chicago bears 24 to 13 at halftime
2	no i meant the san francisco giants baseball team they are playing the cubs	the san francisco giants lost 2 to 0 to the new york yankees last night and their next game is against the chicago cubs at 7 pm eastern tonight. do you want me to remember that the san francisco giants are your favorite baseball team
3	no not really thanks though	no problem. did you want a news update for your other favorite sports teams
4	sure	here is your sports update. the new york...

Figure 1: Example interaction between human end user and conversational AI assistant

Perhaps more importantly, the factors relevant to whether a single dialog turn is "successful" (or "defective") can be quite different from the factors relevant to whether a multi-turn dialog is successful: even human-to-human linguistic communication frequently involves temporary misunderstandings, clarification, rephrasing, etc.—attributes that are indicative of poor interaction quality *only when viewed in isolation*. For example, consider the (synthetic) dialog in Figure 1: Even though the system response in Turn 1 does not satisfy the user's request, the user quickly clarifies, and immediately

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 186–195
July 10-12, 2023 ©2023 Association for Computational Linguistics

receives relevant information. Furthermore, Turn 2's response shows that the user's initial request was based on an incorrect assumption (that a SF Giants game is underway). Despite this, the system provides enough pertinent information to resolve the original request. Viewed as a whole, this is a high-quality dialog.

In this contribution, we present a scalable approach to dialog-level quality estimation based on a new annotation scheme we call Dialog Quality Annotation (DQA). DQA adapts and extends Bodigutla et al.'s (2019b) turn-level Response Quality (RQ) annotation task to the dialog level. Whereas Bodigutla et al. (2019b) obtain dialog-level quality labels via directly soliciting user-satisfaction ratings, DQA uses expert annotators to collect ground-truth labels.

In line with the results of Bodigutla et al. (2019a), we found that aggregations of turn-level signals are indeed predictive of dialog-level ratings. However, we also found that a supervised approach utilizing both dialog-level signals and aggregated turn-level signals achieves superior performance (F1=.81) compared to aggregation of turn-level features alone (F1=.73; similar to the findings of Bodigutla et al. (2019b) for predicting single-turn ratings). These results have implications for the design of multi-turn interaction quality measurement systems, chief among which is that such systems will achieve superior performance if they include both features computed over entire dialogs and features derived from individual turns of a dialog.

Our contributions are summarized as follows:

1. We develop a high-velocity dialog quality annotation (DQA) scheme and use it to generate dialog-level annotations for 3674 dialogs across 11 different domains.

2. We use the annotated data to train a supervised model for predicting binarized dialog-level quality ratings.

3. We conduct experiments and find that our proposed model outperforms baselines in F1 score, and generalizes better to an unseen domain, thus showcasing the value of high-quality dialog-level annotations.

2 Related Work

Existing research on quality metrics for multi-turn human-computer interactions has focused on either task-oriented dialog systems, or open-domain ("chitchat") systems. The present study concerns largely task-oriented use cases, but given the conversational nature of our platform, chitchat also can (and does) occur in dialogs we evaluate.

2.1 TOD Systems

Task-oriented dialog (TOD) systems help humans to achieve concrete tasks via voice or text interaction. For example TOD systems help users book reservations, communicate with customer service systems, or navigate menus. Evaluating the quality of such interactions requires a dataset of TODs annotated with quality scores. A number of TOD datasets have been released publicly (see §4.1 of Sun et al. 2021), but most are designed to evaluate the performance of dialog understanding tasks like Dialog State Tracking, as opposed to the quality of dialogs from the perspective of successful communication. Many such public datasets were created via Wizard-of-Oz experiments, i.e. human-human interactions where one human plays the role of system and the other of user (Eric et al., 2019). Other datasets were collected by first simulating dialog outlines in the form of API sequences and then asking annotators to expand the outlines into natural language dialogs (Rastogi et al., 2020). A recent study annotated TOD datasets with user satisfaction scores by showing dialogs to annotators and asking them to rate for quality (Sun et al., 2021).

Various annotation schemas have been proposed to label the quality of TODs at the turn-level. In Interaction Quality (IQ), raters were asked to rate each turn on a 1-5 scale, taking into consideration the dialog quality so far (Schmitt et al., 2012). To reduce the cognitive load on annotators, Bodigutla et al. (2019b) proposed the Response Quality (RQ) annotation schema. RQ removed the constraint to keep track of the dialog quality so far, but asked annotators to consider if the next user utterance might contain feedback, such as frustration, rephrasing, etc. The RQ scale is: 1=Terrible (fails to understand user's goal), 2=Bad (understands goal but fails to satisfy it in any way), 3=OK (partially satisfies goal), 4=Good (mostly satisfies goal), and 5=Excellent (completely satisfies user's goal). Another recent study (Sun et al., 2021) collected annotations at the dialog level, using a simple (underspecified) 5-point user satisfaction scale.

Various approaches have been explored to train models to estimate task-oriented dialog quality. Earlier approaches used text-based features from dialogs and trained models like SVMs to predict quality scores. More recent approaches use RNNs (sometimes hierarchical) or BERT to encode dialogs and train models to predict turn- and/or dialog-level quality scores. These approaches model the task either as classification (for discrete quality scores) or regression (for quantitative quality scores). Recent research has explored applications of large language models (LLMs) for dialog-based NLU tasks such as intent recognition and dialog state tracking. Such models have been trained using publicly available TOD datasets, e.g. Wu et al. (2020); Peng et al. (2020); Yang et al. (2021). TOD-based LLMs have not been explored as extensively for the purpose of TOD quality estimation, though this is an active area of research for us.

See Deriu et al. 2021 for a survey of approaches to evaluation in TOD systems.

2.2 Open-Domain Dialog Systems

Developing quality metrics for open-domain dialog systems presents different challenges than for TOD systems. In an open-domain dialog, a system can have many relevant responses for a single utterance, and a single dialog could cover multiple unrelated topics. Automated evaluation approaches have explored different aspects of dialog quality such as coherence, informativeness, user engagement (Vakulenko et al., 2018; Zhang et al., 2021; Mehri and Eskénazi, 2020; Ghazarian et al., 2020). Similar to TOD, open-domain dialog evaluation requires high-quality training data. Existing work has used datasets by collecting human judgments (Higashinaka et al., 2014; Cervone and Riccardi, 2020). Another general approach is to use conversations between human users as coherent/positive examples, and then generate negative examples/incoherent dialogs by applying certain perturbations to the coherent dialogues, such as shuffling order or injecting irrelevant utterances into the dialog (Vakulenko et al., 2018; Mesgar et al., 2020; Huang et al., 2020; Zhang et al., 2021). Recent work has considered higher-level semantic perturbations that change the dialog flow more subtly (Ghazarian et al., 2022b).

3 Dialog Quality Annotation

3.1 DQA Workflow

Here we describe the workflow for generating annotations needed to train a supervised dialog quality estimation model. This workflow adapts and extends the related turn-level Response Quality (RQ) workflow of Bodigutla et al. (2019a). We refer to this workflow as "Dialog Quality Annotation" (DQA). DQA is platform- and domain-agnostic, and was designed to support high-velocity annotation.

In each DQA task, a multi-turn dialog is presented in its entirety to an expert data annotator (DA). First, the DA is asked to rate the quality of each turn in the dialog. After each turn has been annotated, the DA then answers questions about the dialog as a whole (overall dialog rating, number of goals, goal completion, goal progression, goal friction, system response coherence, and user's inferred sentiment). DAs assigned quality scores to dialogs using a five-point rating scale. About 20% of dialogs are annotated by two DAs, for quality control monitoring. After the workflow was fully productionized and DAs were calibrated on the annotation task, we have observed weekly inter-rater agreement rates ranging from 79% to 86% (with a difference of one scale point allowed). See Appendix A for further details about the workflow.

Using the DQA workflow, we gathered a dataset of 3569 annotated dialogs (9347 turns from 3233 unique users), of which 714 were held out as a test set to evaluate the performance of baseline methods and trained models. The remaining 2855 annotated dialogs were used to train candidate dialog-level defect detection models. This data was gathered by randomly sampling (de-identified) interactions across 10 different experiences supported by our platform. Our train-test split was stratified by experience, so that each use case appears at a similar rate across train and test sets.

Finally, we gathered 105 additional annotated dialogs (502 total turns) from a use case that does not appear in the training or test data (Shopping product Q&A). These out-of-distribution (OOD) dialogs enable us to more realistically assess how well the resulting model generalizes to patterns unseen during training.

The majority of the data we gathered were from experiences in which the system only has access to information about the target use case. Such traffic is partitioned into discrete user sessions by default, so we considered a "dialog" to just be a single user session. For the OOD traffic, which does not come with pre-defined session boundaries, we used a time-based heuristic where a dialog is considered to be a sequence of utterances from a single user, with no more than 180 seconds of inactivity between turns. In future work, we are exploring model-based methods for dialog segmentation.

3.2 Dialog quality versus dialog attributes

As discussed above, for every dialog, the DAs provide the overall dialog-level rating, salient attributes of the dialog, and the individual turn-level ratings. With these annotations we aim to understand the relationship between salient attributes of a dialog (e.g. goal progression, goal completion, response coherence) and the overall dialog-level ratings. The motivation here is that a robust relationship between objective dialog attributes and dialog ratings would help us to derive human-quality labels from automated methods in the future. While some research exists on the relationship between turn-level and dialog-level quality ratings (Bodigutla et al. 2020), few studies explore the relationship between dialog-level attributes and dialog-level quality ratings (Siro et al. 2022).

In Figure 2 we plot the distribution of dialog-level rating against four salient attributes of the dialog. As expected we can clearly see that dialogs received higher ratings when users successfully completed their goals, system responses were coherent, and users encountered less friction while progressing towards their goals. Further, Table 1 computes the Spearman's ρ correlation between the ratings and attributes. Goal completion was found to have the highest correlation score of .859, while user sentiment had the lowest, at .449. Moreover, user friction encountered had a negative correlation to dialog rating. These observations are intuitive given the dialogs were sampled from mostly task-oriented experiences.

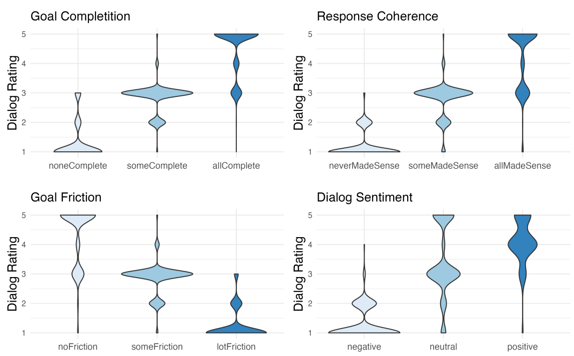

Figure 2: Distribution of dialog ratings with salient attributes of dialog.

4 Dialog Quality Estimation Model

We now describe our dialog quality estimation model (DQM), which leverages the dialog-level annotations described in the previous section.

Table 1: Correlation of dialog rating with salient attributes of the dialog. All correlations in this table are statistically significant at $p < 0.01$.

Attribute	Spearman's ρ
Goal Completion	.859
Response Coherence	.766
Goal Friction	(.807)
User Sentiment	.449

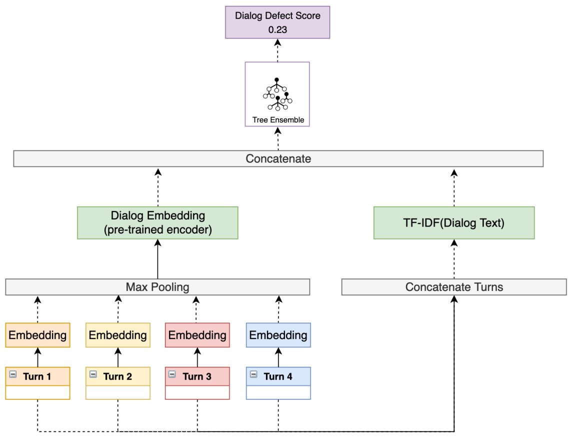

Figure 3: DQM Model Architecture

Figure 3 illustrates the architecture of the model. We first leverage a pre-trained turn-level defect detection model (which is trained on millions of interactions) as a feature extractor using a RoBERTa-IQ-based framework (Gupta et al., 2021). We encode each turn of a dialog as a dense vector. We use a max-pooling operation on the turn-level vectors to obtain a dialog-level representation. Finally, we concatenate this with a bag-of-words representation (TF-IDF over unigrams) of the dialog text. This final dialog-level vector is then fed into a Random Forest Classifier to learn a mapping from dialog-level representation to the binarized defect label in $\{0, 1\}$. We arrived at this setup after experimenting with various text and numeric features, and simpler classification algorithms. We also found more sophisticated models to be less effective with our current dataset, although we plan to revisit more complex architectures as the data grows.

Experimental results comparing the performance of this model against several strong baselines are presented in §5-6.

The pre-trained text encoder used in our model is based upon an internal model that produces turn-level defect (TLD) scores, which are real-valued scores in $[0, 1]$ that can be interpreted as the probability that a given turn is defective from the perspective of the user (see Gupta et al., 2021 for details on the model). TLD scores are derived from a RoBERTa-IQ classifier trained to detect defective turns within a dialog. Although the TLD model does take context into account when scoring interactions, it is explicitly designed to score dialog *turns*, as opposed to entire dialogs.

Our primary question was therefore whether a model trained on the task of dialog-level defect detection outperforms methods that only involve aggregation of turn-level signals. The relevant tradeoff here is that aggregations of TLD scores are cheap and easy to compute, but may suffer from poor accuracy since they were not designed to make predictions about dialogs as a whole.

We hypothesized that a dialog-level statistical model would outperform the TLD-based baselines, in large part because of observed interaction patterns in which the quality of a dialog is not a straightforward combination of the quality of its constituent turns.

5 Experiment Setup

For the purposes of these experiments, we binarized the five-point dialog-level quality labels by assigning dialogs rated 1, 2, or 3 to the defect class, and dialogs rated 4 or 5 to the non-defect class. This follows the approach taken by Gupta et al. (2021) for turn-level response quality prediction, enabling us to frame defect detection as a binary classification task.

We assessed the quality of each estimator by measuring its precision, recall, and F1 score relative to human labels on the held-out test set.

We computed four baseline dialog-quality scores, all of which were derived by aggregating TLD scores across each turn in a dialog. We expected to see a very strong relationship between average TLD and dialog quality score, especially since the TLD model uses information from surrounding turns as features.

These are the baseline methods we computed over the test data used for model evaluation. Each score reduces a sequence of turn-level scores from a dialog into a single value, which represents the dialog-level score.

1. *Mean TLD*: Simple arithmetic mean of the predicted turn-level TLD model scores.

2. *Last-turn TLD score*: Interpret the final turn's TLD score as the dialog-level score. The idea is that recency bias will lead the final turn to have more impact than others in perceived dialog quality.

3. *Union of mean and last-turn TLD*: A dialog is considered defective if either the mean or last-turn TLD score exceeds some threshold (here: 0.5).

4. *Rising linear weights*: Calculate mean TLD score with each turn linearly weighted by its index, so that later turns have higher weights.

Baseline methods required no training process at all, as they consist of arithmetic aggregations of TLD scores, which were already available prior to experimentation. To prepare each dialog for baseline evaluation, we simply computed each aggregation for each dialog. Baseline aggregations were then converted to binary predictions via a threshold: dialogs with scores $\geq .5$ are considered defective; scores $< .5$ are considered non-defective (we found that some use cases achieve higher accuracy with higher thresholds, while others benefit from lower thresholds; here we use the fixed value of .5, as we intend for these methods to be applicable to any supported use case). We scored each dialog in the 714-dialog test set and the 105-dialog OOD test set for each baseline method, and computed performance metrics of interest relative to the human annotations.

To optimize hyperparameters and perform feature selection for our candidate dialog-level defect detection model, we used five-fold cross validation over the training set, selecting the fit that maximized (mean) F1 over the set of hyperparameters and feature subsets considered. The resulting configuration was then trained against the entire 2855-dialog training set. We then used the resulting model to predict defect class (and class probability) over both test sets, computing and recording the same performance metrics of interest.

6 Results

We present the experimental results of the baseline methods and our supervised model for dialog level defect detection (DQM) in Table 2. We describe our observations and inferences from this comparative study in the following section.

6.1 Performance of baselines and DQM

We observed the following regarding the performance of TLD-based baselines and DQM:

1. *Among the TLD-based baselines, the Union of Mean & Last Turn TLD performs best in all scenarios.* However, in absolute terms, the best baseline is not the best performing method for evaluating dialog quality, and only achieves F1 scores of .77 and .51 compared to human annotation on

Table 2: Performance of TLD-based baselines and supervised model

	Multi-domain test set ($n = 714$)			OOD test set ($n = 105$)		
	Precision	Recall	F1-Score	Precision	Recall	F1-score
Mean TLD	**.84**	.54	.66	.39	.77	.52
Last-turn TLD	.83	.68	.75	.47	.23	.31
Union of mean & last-turn TLD	.82	.73	.77	.38	.77	.51
Rising linear weights	.83	.63	.72	.41	.67	.51
DQM	.78	**.83**	**.81**	**.48**	**.80**	**.60**

the multi-domain and OOD data, respectively.

2. *DQM outperforms the best TLD-based baseline in F1 by 4 and 9 percentage points on the multi-domain and OOD test sets, respectively.* Note that the OOD (Shopping) use case was unseen during training, yet the model achieves an out-of-the-box F1 score of .60 in detecting defective OOD dialogs, compared to only .51 for the best baseline.

3. *DQM has a large advantage in recall over baselines, albeit at the cost of reduced precision.*

6.2 Error analysis

We further analyzed the performance of DQM and baseline methods over the test set, splitting the data by various attributes of interest. We made the following inferences on the basis of these analyses:

1. *Performance of TLD-based baselines and DQM as a function of dialog length indicates that the gap widens as dialog length increases.* Baselines perform better for shorter dialogs ($\leq$ 3 turns) and start to drop in performance as dialog length increases, while DQM's performance improves as dialog length increases. This observation likely explains part of the gap between DQM and baselines on OOD data, since these dialogs tend to be much longer than in our multi-domain dataset (mean of 4.78 turns per dialog versus 2.62). Table 3 shows baseline versus DQM performance over the multi-domain test set, split by dialog length.

2. *DQM has an advantage in detecting defective dialogs that contain a small number of fatal turns*, early on or in the middle of the dialog, which create an overall defective experience. In contrast, TLD-based baselines like mean TLD weight each turn equally and often miss such dialogs. See Appendix B.1 for further discussion of this pattern.

3. *Both TLD-based baselines and DQM struggle to differentiate between user query rephrasing, which is typically a defect, and user query refinement, which is typically not a defect* (see Appendix B.2 for examples). User rephrasing happens when a user request is not successful and the user repeats their request with a slightly different surface form. User refinement occurs when a user iteratively refines a successful search by adding or modifying constraints. We observe that TLD-based baselines have a bias towards incorrectly predicting refinements as defects, possibly because it misclassifies them as rephrases. DQM also struggles with this since it uses TLD as input signal. We hypothesize that these biases may be easier to correct by retraining DQM with targeted multi-turn data than by retraining the TLD model, which is primarily trained on single- or few-turn interaction patterns.

Table 3: Performance (ROC-AUC) of TLD-based baselines and supervised model against dialog length on Multi-domain test set ($n = 714$). TLD-U is union of mean and last-turn TLD (the best baseline).

Dialog Length	n	TLD-U	DQM
Short ($\leq$ 3 turns)	535	.76	.79
Medium (4-6 turns)	149	.73	.80
Long ($\geq$ 7 turns)	30	.69	.84

7 Conclusion

In this study, we presented a new dialog-level annotation workflow DQA, which enables high-velocity labelling of multi-turn human-computer interactions. Our approach is similar to Bodigutla et al. (2020), but differs in that we gather labels from expert annotators instead of end users themselves.

We showed that a supervised model trained on DQA annotations outperforms several strong baselines based on aggregating turn-level defect scores. Furthermore, we observed that the model generalizes better to a previously unseen domain. We also found several qualitative patterns of interest, most notably that DQM's advantage over baselines expands as dialog length increases. These findings jointly lend support for an annotation-based approach to estimating multi-turn interaction quality for large-scale dialog systems.

Limitations

Our proposed approach is designed explicitly for evaluation of task-oriented dialog systems, and is hence unlikely to generalize well to chitchat systems. Most traffic to our platform (and our annotation workflows, including DQA) comes in the form of task-oriented interactions. User turns in the traffic we analyze tend to be quite short (usually less than 20 tokens) and direct, so our model is unlikely to perform as well on dialogs driven by long-form user utterances.

Ethical Considerations

We do not envision any ethical concerns with the research presented here. No customer data is released or presented in this paper, and even our internal data sources are fully de-identified and contain no customer Personal Identifiable Information (PII).

Acknowledgments

We wish to thank: Di Wang and Wenbo Yan of Alexa Shopping for providing the OOD test data; the Amazon Data Services team for their work producing annotations; and the metrics team in Alexa for developing the turn-level model that forms the backbone of the dialog-level model presented here. And thanks to the anonymous reviewers, whose feedback helped to clarify and improve this paper.

References

Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019a. Multi-domain conversation quality evaluation via user satisfaction estimation. *3rd Workshop on Conversation AI: Today's Practice and Tomorrow's Potential, 33rd Conference on Neural Information Processing Systems*.

Praveen Kumar Bodigutla, Aditya Tiwari, Josep Valls-Vargas, Lazaros Polymenakos, and Spyros Matsoukas. 2020. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. In *EMNLP 2020*.

Praveen Kumar Bodigutla, Longshaokan Wang, Kate Ridgeway, Joshua Levy, Swanand Joshi, Alborz Geramifard, and Spyros Matsoukas. 2019b. Domain-independent turn-level dialogue quality evaluation via user satisfaction estimation. *arXiv preprint arXiv:1908.07064*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.

Alessandra Cervone and Giuseppe Riccardi. 2020. Is this dialogue coherent? Learning from dialogue acts and entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020*, pages 162–174. Association for Computational Linguistics.

Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In *Proceedings of the ACM Web Conference 2022*, pages 2998–3008.

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

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2021. User response and sentiment prediction for automatic dialogue evaluation. *arXiv preprint arXiv:2111.08808*.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022a. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.05, pages 7789–7796.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022b. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Saurabh Gupta, Xing Fan, Derek Liu, Benjamin Yao, Yuan Ling, Kun Zhou, Tuan-Hung Pham, and Edward Guo. 2021. RoBERTaIQ: An efficient framework for automatic interaction quality estimation of dialogue systems. In *Proceedings of DeMaL, Second International Workshop on Data-Efficient Machine Learning (KDD 2021)*.

Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo. 2014. Evaluating coherence in open domain conversational systems. In *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 130–134. ISCA.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. 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 2020, Online, November 16-20, 2020*, pages 9230–9240. Association for Computational Linguistics.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Carel van Niekerk, Michael Heck, Shutong Feng, and Milica Gašić. 2022. Dialogue evaluation with offline reinforcement learning. *arXiv preprint arXiv:2209.00876*.

Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020*, pages 225–235. Association for Computational Linguistics.

Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. 2020. Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1439–1450. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Schema-guided dialogue state tracking task at dstc8. *arXiv preprint arXiv:2002.01359*.

Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication*, 74:12–36.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU Let's Go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3369–3373.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding user satisfaction with task-oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2499–2506.

Stefan Ultes, Robert ElChab, and Wolfgang Minker. 2014. Application and evaluation of a conditioned hidden markov model for estimating interaction quality of spoken dialogue systems. In *Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice*, pages 303–312. Springer.

Svitlana Vakulenko, Maarten de Rijke, Michael Cochez, Vadim Savenkov, and Axel Polleres. 2018. Measuring semantic coherence of a conversation. In *International Semantic Web Conference*, pages 634–651.

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

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

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. *arXiv preprint arXiv:2004.06871*.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with GPT-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35.16, pages 14230–14238.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. 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*, pages 5676–5689. Association for Computational Linguistics.

A Dialog Quality Annotation Workflow Design

Here are some selected questions for collecting human annotations used in the DQA workflow. The design of this workflow was inspired by See et al. (2019). In each annotation task, a multi-turn dialog is presented to the data annotator (DA) in its entirety. The dialog consists of a sequence of turns. Each turn consists of a User request and a System response.

Turn Level: First, the DA is asked to rate every turn in the dialog.

Provide an overall rating for the System's response in the current turn

- 1-Terrible
- 2-Bad
- 3-Ok
- 4-Good
- 5-Excellent

Dialog Level: Next the DA is asked to answer a series of dialog-level questions to capture the overall rating along with some salient attributes of the dialog.

[User Satisfaction] Rate the overall user satisfaction based on their interaction in the dialog

- 1-Very Dissatisfied
- 2-Dissatisfied
- 3-Normal
- 4-Satisfied
- 5-Very Satisfied

[Goal Completion] How many goals are in the dialog?

- Zero
- One
- Many

[Goal Progression] Did the user make progress towards achieving their goals?

- No Progress
- Some Progress
- Full Progress

[Goal Completion] How many goals did the user complete in the dialog?

- None Completed
- Some Completed
- All Completed

[Goal Friction] Did the user encounter friction trying to achieve their goals in the dialog?

- Lots of Friction
- Some Friction
- No Friction

[Coherence] How often did the System say something which did NOT make sense?

- Never Made Sense
- Some Made Sense
- All Made Sense

[Sentiment] Describe the user's sentiment in the conversation with the System

- Negative
- Neutral
- Positive

B Dialog patterns

In this section, we compare the behavior of the baselines methods and DQM model predictions on specific customer interaction patterns found in multi-turn conversations. Note that the dialog samples in this appendix are synthetic examples fabricated to illustrate common use cases supported by our platform.

B.1 Fatal Turns in a Dialog

Sometimes a dialog can be considered defective based on the interaction in a single turn. We refer to such turns as "fatal turns" (See Turn 2 in Figure 4). This example illustrates that baseline methods which aggregate turn-level information do not adequately capture the non-linear nature of the overall customer satisfaction in a multi-turn dialog. In contrast, we observed that DQM has a higher chance of detecting defective dialogs with fatal turns.

Turn	User utterance	System response	TLD score
1	where is my package	2 packages for your order should arrive today by 10:00 PM. One containing toothbrush. Another package should arrive tomorrow. Containing Smartwatch and Wireless Speaker	0.05
2	what's in my package	I am having trouble accessing your information. Try again later	0.75
3	notify me when they come	Your package containing toothbrush is arriving by 10:00 PM.	0.01
Dialog level ratings			Rating
Mean TLD			0.27
Last-turn TLD			0.01
TLD-U			0.27
DQM			0.80
Human Rating from DQA Workflow			2

Figure 4: Dialog with a fatal turn (Turn 2)

B.2 User Rephrase

Another common dialog pattern is a user-rephrase (See Figure 5). Here the user simply repeats or slightly rephrases their initial request hoping for a better system response. User rephrases in most cases indicate customer friction and often get poor dialog-level ratings. Since the TLD model we use for encoding turns (see § 4) considers context around the current turn, it learns a strong association between user rephrases and defective turns.

Turn	User utterance	System response	TLD score
1	when is the delivery gonna be here	Sorry, I don't have an answer for that	0.99
2	when is my delivery going to get here	Sorry, I don't have an answer for that	0.99
3	when will my package arrive	Your toothbrush, shampoo and 3 more items should arrive by Friday	0.02
Dialog level ratings			Rating
Mean TLD			0.66
Last-turn TLD			0.02
TLD-U			0.66
DQM			0.74
Human Rating from DQA Workflow			2

Figure 5: Dialog with a rephrased user request

B.3 User Refinement

In use cases involving topic exploration or navigation of recommendations, a user progressively adds more constraints to refine their earlier requests (see Figure 6). Unlike the user rephrase pattern described in Section B.2, the previous turns leading to more refinement do not necessarily indicate a unsatisfactory customer experience. A well performing dialog-quality model should learn to distinguish between frictional rephrases and non-frictional refinements. We note that TLD-based baselines have a bias towards incorrectly predicting refinements as defects. As DQM uses TLD as an input signal, DQM too struggles to effectively distinguish between frictional rephrase and non-frictional refinement.

Turn	User utterance	System response	TLD score
1	how much does milk cost	a gallon of milk costs $6	0.90
2	how much does organic milk cost	a gallon of organic milk costs $7	0.20
Dialog level ratings			Rating
Mean TLD			0.55
Last-turn TLD			0.20
TLD-U			0.55
DQM			0.63
Human Rating from DQA Workflow			5

Figure 6: Dialog with a user query refinement