# What Went Wrong?
# Explaining Overall Dialogue Quality through Utterance-Level Impacts

**James D. Finch**[*]
Dept. of Computer Science
Emory University
Atlanta, GA, USA
`jdfinch@emory.edu`

**Sarah E. Finch**[*]
Dept. of Computer Science
Emory University
Atlanta, GA, USA
`sfillwo@emory.edu`

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

## Abstract

Improving user experience of a dialogue system often requires intensive developer effort to read conversation logs, run statistical analyses, and intuit the relative importance of system shortcomings. This paper presents a novel approach to automated analysis of conversation logs that learns the relationship between user-system interactions and overall dialogue quality. Unlike prior work on utterance-level quality prediction, our approach learns the impact of each interaction from the overall user rating without utterance-level annotation, allowing resultant model conclusions to be derived on the basis of empirical evidence and at low cost. Our model identifies interactions that have a strong correlation with the overall dialogue quality in a chatbot setting. Experiments show that the automated analysis from our model agrees with expert judgments, making this work the first to show that such weakly-supervised learning of utterance-level quality prediction is highly achievable.

## 1 Introduction

A typical life cycle of a dialogue system involves many iterative updates where developers improve the system's language understanding capabilities and attempt to increase the overall user engagement. One of the most challenging aspects of executing these updates is to identify characteristics of the dialogue system that are impacting user experience the most. Doing so often involves manually crawling potentially thousands of system logs and designing statistical analyses, both of which are time consuming and unlikely to provide a holistic view of a system's shortcomings.

Inspired by this problem, the presented work investigates the extent to which it is possible to automatically distinguish turns within chat-oriented dialogues that have a negative effect on overall dialogue quality. The interpretation of dialogue quality is especially difficult in the chat-oriented dialogue setting due to its subjective and multi-faceted objectives. System misunderstandings and low user engagement are factors of low quality that are relatively easy to identify, but more subtle factors such as boring responses, awkward topic switches, and individual preferences can also have a substantial effect. Furthermore, the practical value of any approach to estimate the quality of individual dialogue turns is highly sensitive to the cost of collecting relevant data. Chatbots, and the settings they are placed in, can differ drastically in both their topics of conversation and interaction styles. And while conversation-level quality labels can be obtained relatively quickly by asking users to provide a rating at the end of a conversation, collecting data with turn-level labels that adequately characterizes a new chatbot or chat setting is an expensive process.

In this paper we present our dialogue analysis approach, which addresses these challenges by producing quality scores for each utterance in a given conversation dataset using only conversation-level quality ratings. Unlike other work that focuses on utterance-level quality prediction using labeled data, our approach involves training a neural model to learn explicit relationships between utterance-level features and conversation quality without the need for costly utterance-level annotations. We evaluate this approach on two conversation datasets and show high agreement between our model and experts for identifying problematic interactions. By developing an empirical technique that models the relationship between specific interactions and overall conversation quality, our work has the potential to remove much of the human effort and guesswork involved in dialogue system development.

## 2 Related Work

Related work has explored techniques for modelling dialogue quality on both the conversation and utterance level. Sandbank et al. (2018) present

*Proceedings of the Third Workshop on Natural Language Processing for Conversational AI*, pages 93–101
November 10, 2021. ©2021 Association for Computational Linguistics

an approach for classifying low-quality conversations in commercial conversational assistants. Liang et al. (2020) argued against the feasability of conversation-level quality prediction on a Likert-scale and present a pairwise comparison model instead using methods that compensated for the high noise in user scores. Choi et al. (2019) presents methods for both predicting user satisfaction and detecting conversation breakdowns at the turn level. Ghazarian et al. (2020)'s work is similar, predicting utterance-level user engagement.

Ghazarian et al. (2020) and Choi et al. (2019)'s work is similar to ours, as they build models targeted towards utterance-level quality outcomes. However, unlike our approach, these works are reliant on costly turn-level annotations: given conversations annotated for quality on the utterance level, their approach is to train a model that can predict utterance quality on unseen conversations within a similar conversation setting. This strategy incurs a substantial cost whenever the training data needs to be updated to fit a novel conversational setting or chatbot.

To avoid the cost of collecting turn-level labels, our approach is more in line with techniques such as multiple regression analysis, where fitting a model to a dataset is used to explain the relationship between features and some outcome, rather than to predict an outcome for unseen examples. In our case, our model can be fit to any dataset of conversations with conversation-level quality labels in order to estimate the quality impact of each utterance on the overall conversation quality. This approach has a couple advantages over existing work. First, collecting utterance-level annotations in a supervised setting is not necessary for our approach as it was for Choi et al. (2019) and Ghazarian et al. (2020). Second, our model learns empirically-derived relationships between the utterance level and conversation quality, rather than learning to mimic human judgements of utterance level quality irrespective of conversation-level impact. Given the unreliability of human judgements on conversation quality noted in prior work (Liang et al., 2020), it is possible that removing human estimations of the relationship between turn-level and conversation-level quality eliminates a source of bias in the model's training objective. To our knowledge, no previous work has presented an approach for utterance-level quality estimation that does not require data annotated on the turn level.

## 3 Aggregated Regression Analysis

We utilize a neural network model to accomplish this task of utterance-level quality inference. It learns to assign scores to utterances within a dialogue such that it maximizes its ability to produce correct dialogue-level quality scores from aggregation of these utterance scores.

Our Aggregated Regression Analysis (ARA) neural model takes as input the entire sequence of utterances for a single dialogue and outputs the predicted dialogue quality $q$ on a continuous scale. Each utterance is first embedded into a continuous space vector representation, producing a sequence of embeddings $(u_1, ..., u_N)$. The rating $r_i$ and weight $w_i$ of each utterance embedding are then assigned through learned linear transformations of the embedding. The linear transformations are learned independently for the ratings and the weights:

$$r_i = u_i v_r^T + b_r \tag{1}$$

$$w_i = \sigma(u_i v_w^T + b_w) \tag{2}$$

where parameters $v_r$, $b_r$ and $v_w$, $b_w$ are learned weights and biases for the rating and weight calculations, respectively.

The final dialogue quality prediction $q$ is then calculated as the weighted sum of the utterance ratings as follows, where $N$ is the total number of utterances in a given dialogue:

$$q = \frac{\sum\limits_{i=0}^{N} r_i w_i}{\sum\limits_{i=0}^{N} w_i} \tag{3}$$

We also extend this approach to capture inter-utterance interactions by including an utterance contextualization step. The original utterance embeddings $(u_1, ..., u_N)$ are passed through a contextualization layer to generate a contextualized utterance embedding for each utterance $i$ as $h_i$. The resultant rating and weight calculations follow Equations 1 and 2 with $u_i$ replaced by $h_i$. This final architecture is shown in Figure 1.

Although our task is to quantify the relationship between individual utterances and conversation quality, our model is formulated as a conversation-level quality predictor in order to fit to datasets where conversation quality labels are available. Since the top layers of our architecture enforce

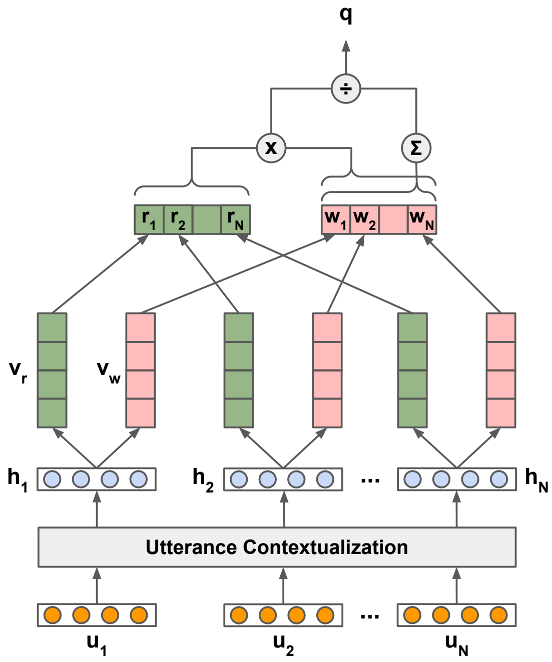

Figure 1: Model architecture for predicting dialogue quality by aggregating predicted utterance ratings and weights with contextualization.

this conversation-level prediction to be constructed from individual ratings and weights of each utterance in the conversation, this conversation quality prediction architecture can be applied to the model analysis task that is the focus of this paper. After fitting this model to a conversation dataset with quality labels, utterance-level effects on quality can be inferred by extracting the ratings and weights of each utterance from the model's intermediate output. The final inferred impact score $s_i$ of each utterance $u_i$ on the conversation quality is then simply given as a product of the rating and weight:

$$s_i = r_i * w_i \tag{4}$$

## 4 Experiment

### 4.1 Data

Our task requires dialogue datasets that contain dialogue-level quality ratings. However, to measure the success of inferring utterance-level impacts from these dialogue-level quality ratings, we need to evaluate any approach on this task against dialogues that also contain utterance-level ratings. To our knowledge, there are two publically available datasets that provide both dialogue-level and utterance-level quality ratings: Amazon's Topical-Chat (Gopalakrishnan et al., 2019) and the First Conversational Intelligence Challenge (ConvAI) (Burtsev et al., 2018).

Although Topical-Chat obtained human annotations on utterance quality, there are several limitations to these annotations that make them unsuitable as an evaluation dataset for our task. For one, the quality of the utterance-level annotations is questionable. Through a preliminary analysis of the dataset, we observed many cases where we could not justify the human annotator's ratings. In fact, 99% of the utterance-level annotations received a rating of 4 or 5, and annotators seemed likely to give such a rating with no regard to the coherence of the utterance in its context. Additionally, even if the annotation reliability was not in question, this dramatic bias towards a small subset of the rating classes also leads to the dataset containing a substantial lack of variety in utterance quality, which would make this dataset uninteresting as an evaluation method. For these reasons, we choose to exclude Topical-Chat from our evaluation.

On the other hand, ConvAI contains more distributed utterance-level ratings that did not raise the same quality concerns upon our preliminary analysis. As a result, we include ConvAI in the evaluation of this task. In addition, we augmented a subset of the dialogues collected through our participation in the 2019 Alexa Prize with utterance-level annotations. We use this augmented dataset for evaluation as well. More details on these included datasets are provided next.

**ConvAI** ConvAI was held as a Competition Workshop at NIPS 2017, where 10 teams submitted bots that were able to hold conversations with humans on short text news articles. During evaluation of the bots, humans provided a dialogue quality rating at the conclusion of their conversation, and also had the option of rating each of their conversational partner's turns using an online good/bad indicator. We heldout 200 conversations from the overall dataset as development and test splits. Table 1 shows statistics of the remaining ConvAI conversations ($N$=2,459) used for training and utterance-level evaluation.

**AP19** AP19 consists of dialogues that we collected during the 2019 Amazon Alexa Prize Competition (Gabriel et al., 2020). These dialogues occurred between Amazon Alexa users and one socialbot from the 2019 competition. This socialbot

| Rating | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Dialogues** | 1100 | 537 | 345 | 308 | 169 |
| **Proportion** | 44.7 | 21.8 | 14.0 | 12.5 | 6.9 |
| **Avg. Turns** | 5.99 | 12.06 | 12.98 | 12.38 | 13.05 |
| **Good Turns** | 406 | 726 | 720 | 1262 | 783 |
| **Bad Turns** | 1371 | 1168 | 623 | 320 | 111 |

Table 1: Statistics for the ConvAI dataset.

was developed to hold in-depth chat-oriented conversations on a wide variety of topics with users, including sports, pets, work, and family. Dialogues were rated by the user at the conclusion of the dialogue on a voluntary basis by providing a numeric score between one and five in response to the question of how much they would want to talk to this socialbot again.

The AP19 dataset was collected between March and July of 2020. Dialogues with less than 5 utterances were excluded because we observed that such dialogues frequently occurred due to unintentional invocation of the Alexa Prize skill. We heldout 4873 conversations as development and test splits each. Table 2 shows statistics of the remaining AP19 conversations ($N$=38,693) used for training and utterance-level evaluation.

| Rating | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Dialogues** | 4785 | 4534 | 5965 | 8385 | 15024 |
| **Proportion** | 12.4 | 11.7 | 15.4 | 21.7 | 38.8 |
| **Avg. Turns** | 26.04 | 30.22 | 32.80 | 36.57 | 34.98 |

Table 2: Statistics for the AP19 dataset.

### 4.2 Models

We trained 3 variants of the ARA model presented in Section 3, including both the non-contextualized base version and two extensions using different contextualization methods:

**Non-Contextualized** This is the base ARA model from Section 3 that does not utilize contextualization of the utterance embeddings.

**Order Driven Contextualization** ARA-O extends the base model by using a bidirectional LSTM layer in order to target the importance of utterance order when determining utterance quality.

**Attention Based Contextualization** ARA-A extends the base model by using a self-attention layer to incorporate long-range cross-utterance relationships when determining utterance quality.

We also include an additional baseline model:

**Non-Aggregated Regression Analysis** NARA is trained on the task of directly predicting utterance scores, instead of a final dialogue score. It employs a heuristic for obtaining target scores for utterances, where each utterance score is equal to the dialogue score in which it occurs. This problem is treated as a sequence regression task, where the model takes as input a sequence of utterances for a given conversation and the utterances are first contextualized using a bidirectional LSTM layer before being passed through a fully connected layer to output the predicted score for each.

### 4.3 Model Configurations

Utterances are embedded using the DistilBERT version of Sentence-BERT (SBERT) by Reimers and Gurevych (2019). SBERT is a sentence encoder utilizing a siamese neural network architecture and BERT-based embeddings that has been shown to outperform other methods of encoding sentences on a variety of downstream NLP tasks.

For all ARA-derivative models, hyperparameters were chosen such that they yielded the best-performing models at predicting conversation quality according to Pearson's correlation on a development set of the data. For NARA, the decision was based on performance on utterance-level quality prediction instead. Details on the final configurations are provided in Appendix A.

### 4.4 Dialogue Quality Regression Results

Table 3 shows the performance of our models on predicting overall dialogue quality. Previous works have shown the difficulty of this task, the noisiness of user ratings, and the low agreement between independent human annotators on the same conversations (Liang et al., 2020; Finch and Choi, 2020). As noted in Section 4.3, we used the development performance for hyperparameter tuning only.

| | **AP19** | | **ConvAI** | |
|---|---|---|---|---|
| **Model** | **Dev** | **Test** | **Dev** | **Test** |
| **ARA** | 0.34 | 0.33 | 0.29 | 0.27 |
| **ARA-O** | 0.37 | 0.36 | 0.44 | 0.39 |
| **ARA-A** | 0.36 | 0.35 | 0.34 | 0.25 |
| **NARA** | 0.36 | 0.35 | 0.40 | 0.20 |

Table 3: Pearson's $r$ correlations achieved by each model on the dev/test data.

## 5 Evaluation

Evaluating our model is challenging because it requires human judgements about the magnitude and direction of effect of each utterance on the quality of the conversation it is a part of. Annotating quality on an utterance level is already a difficult task with much subjectivity involved (Higashinaka et al., 2016). Furthermore, asking humans to annotate quality using real numbered values results in arbitrary judgements of magnitude, further complicating a direct evaluation of our task. Nevertheless, a model that appropriately assigns utterance scores in a way that explains each utterance's relative impact to the conversation quality should agree substantially with judgements of human experts.

We address these challenges by presenting two evaluation procedures that approximate the ground truth of our task formulation while framing all human judgements as non-arbitrary decisions with high inter-annotator agreement. These evaluations focus especially on utterances that have a substantial negative effect on conversation quality, since this case is most in line with the main motivations of our work to automatically identify problematic interactions.

### 5.1 Human-Led Evaluation

Our first evaluation is motivated by the idea that a good model should score interactions with well-known issues lower than most other interactions. To measure this, human annotators provide binary judgements on the quality of each utterance. Given these human annotations, we evaluate our model by treating the binary utterance-level annotations as a partial ordering of the impact of each utterance on the conversation rating against which we can measure the agreement of the model's full ranking of utterances composed of their assigned scores. This agreement is represented by the C-Index as used by Steck et al. (2008), which measures the extent to which pairs of items in a predicted ranking agree with some ground-truth partial ranking.

**ConvAI** The ConvAI dataset was released with utterance-level ratings by human annotators. Human users provided a binary rating to utterances indicating whether it was good or bad on a voluntary basis. Thus, we use these ratings provided by the human users for the automatic evaluation of our proposed approach on the ConvAI dataset.

**AP19** Although AP19 contains dialogue-level quality ratings from human users, it does not contain such ratings for utterances. To this end, we construct an evaluation dataset by asking experts to pick out issues within each dialogue that are likely to cause a lower user rating. We specifically investigate two types of issues: system misunderstandings and user dissatisfaction. These two issue types are prevalent in our AP19 data and are a frequent cause of conversation quality degradation.

Two of the authors annotated issues in these two categories on 100 randomly sampled dialogues from our AP19 dataset. Sampling was restricted to only select dialogues with less than a 5 quality rating, since high-rated conversations often have no major problems. The following guidelines were used to annotate each dialogue:

1. Out of the system utterances that signaled a misunderstanding, mark the one that was most likely to negatively impact the user rating.
2. Out of the user utterances that signaled dissatisfaction, mark the one that most strongly indicated that the user was dissatisfied.
3. If the severity of two or more system misunderstandings or user dissatisfaction signals cannot be discriminated, mark all of them.

All conversations in the sample were doubly annotated, with an interannotator agreement (Cohen's kappa) of 0.674 For evaluation purposes, we ignore utterances that the annotators did not agree on, with the final evaluation containing 158 issue utterances and 3020 non-issue utterances.

#### 5.1.1 Human-Led Evaluation Results

| Model | AP19 | ConvAI |
|---|---|---|
| **NARA** | 0.760 | 0.739 |
| **ARA** | 0.712 | 0.574 |
| **ARA-O** | 0.807 | 0.728 |
| **ARA-A** | 0.638 | 0.599 |

Table 4: Average model performance (C-Index) on ranking utterances with respect to human annotations.

Table 4 shows the performance of our proposed approach in predicting the correct quality ranking of utterances. These results are based on the average performance across three independent iterations of each model fitted on the datasets. Across all versions of the evaluation, our proposed approach using the order-driven contextualization (ARA-O) displays strong performance, achieving a C-Index of 0.807 on AP19 and 0.728 on ConvAI. For ConvAI, however, ARA-O is slightly outperformed by the NARA baseline. For AP19, the attention-based contextualization (ARA-A) configuration performs much worse, producing performance that is lower than either of the baselines, although it outperforms the base ARA for ConvAI. We observe that ARA-A is prone to significant overfitting which may account for this poor performance.

We also investigated the transferability of the best-performing model trained on one dataset (AP19) to another (ConvAI) shown in Table 5. On its own, the model performs poorly on the out-of-domain dialogues which we expected. When fine-tuned on ConvAI, its performance is comparable to a model trained only on the ConvAI dataset.

| | **ConvAI** |
|---|---|
| **AP19 ARA-O** | 0.506 |
| **+ Finetune** | 0.724 |

Table 5: Performance of model trained on AP19 when applied to ConvAI.

We include the C-Index scores of the annotators against each other (Table 6) in order to provide a strong comparison for our models. Although the difference between our best model and human performance is not too great, there is still a small gap in performance.

| | **Annot. 1** | **Annot. 2** |
|---|---|---|
| **Annot. 1** | - | 0.823 |
| **Annot. 2** | 0.853 | - |

Table 6: Human annotator's agreement (C-Index) on utterance-level issue scores for AP19.

## 5.2 Model-Led Evaluation

A major limitation of our human-led evaluation is the restriction to only two kinds of issues. Although humans seem capable of judging when well-known problems like misunderstandings and dissatisfaction impact the dialogue quality, there are a number of non-obvious factors that contribute to the user's experience as well. Data-driven models have a greater potential than humans to uncover these factors, since they can make inferences based on information aggregated across thousands of conversations rather than on biased intuition. However, given that such a model outputs some estimation of an utterance's impact on conversation quality, that output should be verifiable by a human and agree substantially with expert judgement.

To account for a broader spectrum of factors that may affect conversation quality, we conducted an additional evaluation that asks experts to judge pairs of interactions, where one interaction was assigned a low score by the model and one was assigned a high score. We choose the 5th percentile as a cutoff to distinguish issues (low-scored) from non-issues (high-scored), and randomly sample one utterance from each side of the distribution to construct pairs. Sampling is not done at uniform random, since we observed in a pilot evaluation that this results in many similar samples that are often paraphrases of one another. Instead, we run $k$-means clustering on the pretrained SBERT embeddings of the bottom 5%-scored utterances and restrict our sampling to the $k$ utterances closest to the centroid of each cluster. We chose $k$ to be 1% of the total number of issue utterances being clustered, $k = 683$. This procedure ensures a high degree of variety among chosen samples, increasing the robustness of the evaluation.

Each sampled issue from the clustering procedure is paired with a non-issue drawn uniform-randomly from the upper 95%-scored utterances. Each pair of utterances is then presented to the human judge with random intra-pair order, so the judge is blind to which utterance was scored lower by the model. We allow judges to see two preceding utterances and one following utterances from the full conversation when viewing each example. Providing a context window of this size was done in order to focus the judgement on a specific interaction of the dialogue, while still providing sufficient background to interpret the sampled utterance correctly. Given expert judgements on these pairs, model accuracy is calculated as the proportion of times expert judgements agreed with model score assignments. We retrieve 300 pairs for evaluation from our AP19 dataset and two of the authors performed the evaluation.

### 5.2.1 Model-Led Evaluation Results

We conducted the Model-Led Evaluation on ARA-O for our AP19 dataset, since it outperformed all other models in the initial Human-Led Evaluation. Table 7 presents the model's ability to discriminate quality ratings between utterances. Our model is able to achieve an accuracy of 0.775 on average.

## 6 Error Analysis

We conducted an error analysis to determine the frequently occurring mistakes of the proposed approach. To do this, we manually reviewed the 70 utterance pairs and their contexts that resulted in disagreement between the system and at least one

| | Accuracy |
|---|---|
| **Annot. 1** | 0.77 |
| **Annot. 2** | 0.78 |
| **Average** | 0.775 |

Table 7: Model accuracy at identifying low-quality utterances. **Annot #** is relative to individual annotators' selections. **Average** takes the average of the annotators' accuracy results.

| Mistake Type | Frequency | Percentage |
|---|---|---|
| Both Issues | 37 | 53% |
| Both Non-issues | 24 | 34% |
| Flipped Assignment | 9 | 13% |

Table 8: Distribution of the three mistake types.

annotator in the System-Led Evaluation. We created 3 categories that captured the underlying problems that caused the disagreement. *Both Issues* contains cases where the non-issue selection was as problematic as its issue selection. *Both Non-issues* represents cases where the issue selection was not interpretable as a problematic conversation exchange. *Flipped Assignment* captures the cases where the non-issue selection was more problematic than the issue selection. Table 8 illustrates the distribution of identified mistake categories across the 70 pairs, with examples shown in Table 10.

| Dialogue Pattern | Issue Prop. | Nonissue Prop. |
|---|---|---|
| Stop Request | 20% | 7% |
| Irrelevant Topic | 16% | 6% |
| Topic Switch Request | 13% | 3% |
| Unengaging Handoff | 11% | 3% |

Table 9: Distribution of the dialogue patterns within system-selected issue contexts and nonissue contexts.

Upon closer analysis of these disagreed pairs, we observed some dialogue patterns that had high correlation with the system assigning a low score to an utterance as shown in Table 9. Using a context window of size 2 around the selected issue, we observed the following patterns:

**Stop Request** the selected issue was within one turn of the user requesting the conversation to be over.

**Irrelevant Topic** the selected issue occurred near the user sharing that they did not have a particular trait (e.g. not in school, no job, no pets, etc.).

**Topic Switch Request** the selected issue was within one turn of the user requesting a new topic.

**Unengaging Handoff** the selected issue was within one turn of the system transitioning out of one topic using one of three unengaging statements, such as *Ok well I find it inspiring that we have so many different types of music to listen to*.

The frequency of these dialogue patterns is shown in Table 9 and examples illustrating these dialogue patterns can be found in Table 10.

These dialogue patterns were present in many of the pairs in the *Both Non-issues* class. In isolation, the instances of these dialogue patterns do not seem indicative of a problematic conversational interaction; however, when taken as an emergent pattern, they begin to illuminate types of interactions that may be subtly indicative of poor user experience. For instance, there would be no reason to request a new topic or to end the conversation if the user was enjoying themselves, thus these actions are a good indication of the user being dissatisfied. Similarly, our socialbot was better equipped to talk about certain life traits, such as being in school or having a pet, and the analogous conversation for those users who did not fit those categories were not as strong. It could be the case that the model's selection of such utterances was not incorrect and rather it was reaching appropriate conclusions based on a large volume of information it was working with, whereas the human annotators were unaware of this, although it is impossible to verify this through our evaluation procedure.

On the other hand, for the *Stop Request* category, it could also be the case that the model is overrelying on the stop utterances as a negative signal. Knowing the end of the conversation provides the model with the opportunity to make any last-second corrections to achieve the dialogue quality it needs to assign for the current conversation.

## 7 Discussion

By achieving a C-Index score of 0.80 and 0.72 on the AP19 and ConvAI datasets respectively, our approach demonstrates the feasability of inferring utterance-level quality information from aggregation over a dataset of rated conversations. Since C-Index is a generalization of the AUC curve, random decision-making would result in a performance of 0.5, thus we obtain improvements of over 20% against random.

The success of our proposed approach on two datasets also demonstrates its applicability across different dialogue settings. A key difference be-

**Example of system-chosen issue utterance:**

**System:** Oh, that is unfortunate, you should definitely try it sometime, V R is slowly becoming the future of gaming. Another emerging area is augmented reality, or A R . what do you think of augmented reality?
**User:** i don't know what that is
**System:** *You don't have any ideas? That is ok. It is still a relatively new thing, so it may be hard to form an opinion. If you were to play a mobile game using A R, what would make it the most fun, do you think?*
**User:** what is a r.

**Example of system-chosen issue utterance with *Topic Switch Request*:**

**System:** Yeah. You know I hear the grocery stores in affected areas have lots of empty shelves right now. I'm not sure how necessary it is, but have you been stocking up on supplies?
**User:** yeah
**System:** *I guess you can never be too prepared. You know, sometimes I think, if everyone just stayed home and avoided public areas for a couple weeks, the whole thing could just blow over. But realistically, I think slowing the virus down so that everyone doesn't get sick at once is the best we can hope for.*
**User:** let's talk about the drivers

**Example of system-chosen non-issue utterance to which annotators disagreed:**

**System:** Yeah, that is true. There are probably many difficult pieces in the puzzle of teleportation, but we can leave that up to the scientists. I cannot seem to remember any movies that involve teleportation, can you?
**User:** i don't know
**System:** *Good one. I will have to remember that. Ok. tell me, what was the last movie you liked?*
**User:** i like joker

Table 10: Examples illustrating highlights and shortcomings of the proposed approach, where the utterances assigned a low quality are indicated by italics. These examples are similar to interactions seen in the Alexa Prize logs (conversations have been modified to protect customer privacy).

tween the AP19 and ConvAI datasets is their rating distributions: ConvAI was more right-skewed whereas AP19 was more left-skewed. In addition, the ConvAI dataset employs a text-based interface, includes both human-machine and human-human dialogues, and the dialogues were less socially-oriented. This suggests that our approach to utterance-level explainability is not constrained to only one dialogue system configuration.

It is worth noting that the NARA baseline slightly outperforms the presented approach on the ConvAI dataset. This suggests that parsing through all of the variability present in noisy quality ratings across users to determine utterance-level correlations is potentially challenging when given a small dataset. It may be more suitable to use the conversation rating as a heuristic utterance-level score target in these cases. However, the difference between the proposed approach and this baseline is not too great, so it is difficult to make a strong conclusion.

## 8 Conclusion

Our approach and evaluation results demonstrate the feasibility of an automatic approach for dialogue quality analysis. By training a model to learn the relationship between utterance-level features and conversation-level outcomes, it is possible to obtain an empirically-derived ranking of interactions based on whether they relate to positive or negative conversation quality. This work has obvious applications in dialogue system development, mitigating the high developer effort involved in manual analysis and human bias in determining the relative importance of system issues.

## References

Mikhail Burtsev, Varvara Logacheva, Valentin Malykh, Iulian Vlad Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, and Yoshua Bengio. 2018. The first conversational intelligence challenge. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 25–46. Springer.

Jason Ingyu Choi, Ali Ahmadvand, and Eugene Agichtein. 2019. Offline and Online Satisfaction Prediction in Open-Domain Conversational Systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 1281–1290, Beijing, China. Association for Computing Machinery.

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

Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, Shui Hu, Karthik Gopalakrishnan, Seokhwan Kim, Lauren

Stubel, Kate Bland, Arindam Mandal, and Dilek Hakkani-Tür. 2020. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Alexa Prize 3 Proceedings*.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive Engagement: An Efficient Metric For Automatic Evaluation of Open-Domain Dialogue Systems. *arXiv:1911.01456 [cs]*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3146–3150.

Weixin Liang, James Zou, and Zhou Yu. 2020. Beyond User Self-Reported Likert Scale Ratings: A Comparison Model for Automatic Dialog Evaluation. *ACL 2020*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3973–3983.

Tommy Sandbank, Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, John Richards, and David Piorkowski. 2018. Detecting Egregious Conversations between Customers and Virtual Agents. *Proceedings of NAACL-HLT*.

Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar. 2008. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pages 1209–1216.

## A Data Preprocessing and Model Hyperparameters

**AP19** All utterances in the AP19 dataset were preprocessed by removing all punctuation. All models were trained using the Adam optimizer with a learning rate of 0.00001 and Mean Squared Error loss function. All models had a dropout layer of 0.1 applied to the utterance embeddings. The *ARA-O* and *NARA* models had a hidden dimension of 200. The *ARA-A* had 1 layer with 1 head, hidden dimension 200, and dropout of 0.1 in the Transformer layer.

**ConvAI** All utterances in the ConvAI dataset were preprocessed by converting all symbols to a language representation (e.g. **>:(** to 'angry face', **lol** to 'laughing out loud', and so on) and then removing all remaining punctuation. In addition, multiple sequential utterances by the same speaker in a conversation were concatenated together into a single response.

All models were trained with the Adam optimizer with a learning rate of 0.00001 and Mean Squared Error loss function. All models had a dropout layer applied to the utterance embeddings (*ARA*: 0.2, *Others*: 0.1). The *ARA-O* and *NARA* models had a hidden dimension of 200. The *ARA-A* had 1 layer with 1 head, hidden dimension 100, and dropout of 0.1 in the Transformer layer.