

Joint Turn and Dialogue level User Satisfaction Estimation on Multi-Domain Conversations

Praveen Kumar Bodigutla

pbodigutla@linkedin.com*

Aditya Tiwari

aditiwar@amazon.com

Josep Valls Vargas

jvalls@amazon.com

Lazaros Polymenakos

polym1@amazon.com

Spyros Matsoukas

matsouka@amazon.com

Abstract

Dialogue level quality estimation is vital for optimizing data driven dialogue management. Current automated methods to estimate turn and dialogue level user satisfaction employ hand-crafted features and rely on complex annotation schemes, which reduce the generalizability of the trained models. We propose a novel user satisfaction estimation approach which minimizes an adaptive multi-task loss function in order to jointly predict turn-level Response Quality labels provided by experts and explicit dialogue-level ratings provided by end users. The proposed BiLSTM based deep neural net model automatically weighs each turn's contribution towards the estimated dialogue-level rating, implicitly encodes temporal dependencies, and removes the need to hand-craft features.

On dialogues sampled from 28 Alexa domains, two dialogue systems and three user groups, the joint dialogue-level satisfaction estimation model achieved up to an absolute 27% (0.43 \rightarrow 0.70) and 7% (0.63 \rightarrow 0.70) improvement in linear correlation performance over baseline deep neural net and benchmark Gradient boosting regression models, respectively.

1 Introduction

Automatic turn and dialogue level quality evaluation of end user interactions with Spoken Dialogue Systems (SDS) is vital for identifying problematic conversations and for optimizing dialogue policy using a data driven approach, such as reinforcement learning. One of the main requirements to designing data-driven policies is to automatically and accurately measure the success of an interaction. Automated dialogue quality estimation approaches, such as Interaction Quality (IQ) (Schmitt et al., 2012) and recently Response Quality (RQ)

(Bodigutla et al., 2019a) were proposed to capture satisfaction at turn level from an end user perspective. Automated models to estimate IQ (Ultes et al., 2014; Schmitt et al., 2011; Asri et al., 2014) used a variety of features derived from the dialogue-turn, dialogue history, and output from three Spoken Language Understanding (SLU) components, namely: Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), and the dialogue manager. RQ prediction models (Bodigutla et al., 2019a) further extended the feature sets with features derived from the dialogue-context, aggregate popularity and diversity of topics discussed within a dialogue-session.

Using automatically computed diverse feature sets and expert ratings to annotate turns overcame limitations suffered by earlier approaches to measure dialogue quality at turn-level, such as using sparse sentiment signal (Shi and Yu, 2018), intrusive solicitation of user feedback after each turn, and using manual feature extraction process to estimate turn-level ratings (Engelbrecht et al., 2009; Higashinaka et al., 2010).

For predicting user satisfaction at dialogue-level, IQ estimation approach was shown to generalize to dialogues from different domains (Schmitt and Ultes, 2015). Using annotated user satisfaction ratings to estimate dialogue-level quality, overcame the limitation with using task success (Schatzmann et al., 2007) as dialogue evaluation criteria. Task success metric does not capture frustration caused in intermediate turns and assumes the end user goal is known in advance. However, IQ annotation approach to rate each turn incrementally, lowered Inter Annotator Agreement (IAA) for multi-domain dialogues (Bodigutla et al., 2019b). Multi-domain dialogues are conversations that span multiple domains (Table 1) in a single dialogue-session. On the contrary, RQ ratings were provided for each turn independently and were shown to be highly con-

* Currently at LinkedIn, but did this work at Amazon.

sistent, generalizable to multiple-domain conversations and were highly correlated with turn-level explicit user satisfaction ratings (Bodigutla et al., 2019b). Furthermore, using predicted turn-level RQ ratings as features, end-user explicit dialogue-level ratings for complex multi-domain conversations were accurately predicted across dialogues from both new and seasoned user groups (Bodigutla et al., 2019b). Earlier widely used approach, such as PARADISE (Walker et al., 2000), where the model is trained using noisy end dialogue ratings provided by users, did not generalize to diverse user population (Deriu et al., 2019).

Despite generalizing to different user groups and domains, both turn and dialogue level quality estimation models trained using annotated RQ ratings (Bodigutla et al., 2019a,b) used automated, yet hand-crafted features. Modern day SDS support interoperability between different dialogue systems, such as “pipeline based modular” and “end-to-end neural” dialogue systems. Hand-crafted features designed based on one system are not guaranteed to generalize to dialogues on a new system.

RQ based dialogue-level satisfaction estimation models (Bodigutla et al., 2019b) did not factor in noise in explicit user ratings and used average estimated turn-level RQ ratings as a feature to train the model. Each turn’s success or failure was assumed to have an equal contribution to the overall dialogue rating. However, a user might be dissatisfied even if most of the turns in the dialogue were successful (example in Appendix Table 8).

The LSTM (Hochreiter and Schmidhuber, 1997) based IQ estimation approaches (Pragst et al., 2017; Rach et al., 2017) were shown to encode temporal dependencies between turns implicitly. Most recently, BiLSTMs (Bi-directional LSTMs) with self-attention mechanism (Ultes, 2019), which used only turn-level features achieved best performing IQ estimation performance.

In order to address the aforementioned limitations with using hand-crafted features, we propose a LSTM (Hochreiter and Schmidhuber, 1997) based turn-level RQ estimation model, which implicitly encodes temporal dependencies and removes hand-crafting of turn and temporal features. Along with turn-level features that are not dialogue-system or user group specific, we use features derived from pre-trained Universal Sentence Encoder (USE) embeddings (Cer et al., 2018) of an utterance and system response texts to train the model.

User utterance {Domain}	System Response
Movies playing near me {MovieBooking}	Avengers is playing at AMC
What are the showtimes? {MovieBooking}	4 p.m and 7 p.m
Two tickets for 7 p.m show {MovieBooking}	Booking 2 tickets for 7 p.m ...
Is it raining today? {Weather}	Yes there is a 60% chance of rain
Book cab to the theater at 6 p.m {CabBooking}	Ride for 2 costs \$20, should I book?
Yes {CabBooking}	Sure, booked a cab for 2 to AMC
Cancel 7 p.m appointment {Calendar}	Canceled your 7 p.m park visit

Table 1: Example multi-domain dialogue a user has with the conversation agent to plan his/her evening. The conversation spans 4 domains in a single dialogue session.

Pre-trained sentence representations provided by USE Transformer model achieved excellent results on semantic relatedness and textual similarity tasks (Perone et al., 2018).

Using an adaptive multi-task loss weighting technique (Kendall et al., 2017) and attention (Vaswani et al., 2017) over predicted turn-level ratings, we further extend the turn-level model to design a novel BiLSTM (Graves et al., 2013) based joint turn and dialogue-level quality estimation model. To test the generalization performance of the proposed approaches, we estimate turn and dialogue-level ratings on multi-turn¹ multi-domain conversations sampled from three user groups, spanning 28 domains (e.g., *Music, Weather, Movie & Restaurant Booking*) across two different dialogue systems.

To the best of our knowledge, this is the first attempt to leverage noise adaptive multi-task deep learning approach to jointly estimate annotated turn-level RQ and user provided dialogue level ratings for multi-domain conversations from multiple user groups and dialogue systems.

The outline of the paper is as follows: Section 2 discusses the choice of RQ annotation. Section 3 & 4 presents the novel approaches to estimate turn and dialogue level quality ratings. Section 5 summarizes the turn and dialogue level data and presents our experimental setup. Section 6 provides an empirical study of the models’ performance. Section 7 concludes.

2 Response Quality for Turn and Dialogue level Quality Estimation

Interaction Quality (IQ) (Schmitt et al., 2012) and Task Success (TS) (Schatzmann et al., 2007) measures require an annotator to accurately determine the task that the user is aiming to accomplish through a dialogue, which is non-trivial for multi-domain conversations (Bodigutla et al., 2019b).

¹ In single-turn conversations the entire context is expected to be present in the same turn. In multi-turn case context from previous turns is carried to address user’s current request.

Both IQ and RQ (Bodigutla et al., 2019a) require annotators to rate each turn on a discrete five point scale (RQ rating scale in Appendix Table 9). Unlike IQ, RQ annotators need not keep track of dialogue progression so far to rate an individual turn. Due to the simplicity of annotation scheme, multi-domain generalizability and applicability to dialogue level satisfaction estimation (Bodigutla et al., 2019b), we chose turn-level RQ annotation scheme. Similar to (Bodigutla et al., 2019b), the dialogue-level ratings are directly obtained from end users who interacted with different dialogue systems. Unlike TS metric, which does not capture user’s dissatisfaction in intermediate turns, dialogue-level satisfaction ratings holistically capture the overall satisfaction of an end user’s interaction with SDS.

3 Turn-level Dialogue Quality Estimation

In this section we discuss previous turn-level satisfaction estimation models trained using RQ ratings, their limitations and our approach to overcome them.

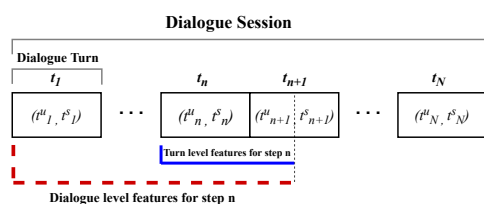


Figure 1: Dialogue and turn definitions for estimating user satisfaction rating on turn t_n (Bodigutla et al., 2019a). The solid blue and dotted red lines indicate the context used for generating turn and dialogue level features respectively.

Similar to Bodigutla et al. (2019a), we define a dialogue turn at time n as $t_n = (t_n^u, t_n^s)$, where t_n^u and t_n^s represent the user request and system response on turn n respectively (Figure 1). A dialogue session of N turns is defined as $(t_1:t_N)$. In experiments conducted by Bodigutla et al. (2019a), Gradient Boosting Regression (Friedman, 2001) model gave the best turn-level RQ prediction performance. Features used to train the model were derived from current turn (t_n), dialogue history ($t_{1:n-1}$) and next turn’s user request (t_{n+1}^u). In addition to deriving domain-independent features from three SLU components, namely Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), and the dialogue manager, five new feature sets were introduced by the authors to improve the performance of the turn-level satisfaction estimation model.

Features used in the model were automatically

computed, yet they were carefully hand-engineered (See Appendix Table 11). Features were hand-crafted to identify and rank factors contributing to the predicted satisfaction rating, but these features do not generalize easily to different dialogue systems. Introduced originally by authors of RQ, “un-actionable request” feature was computed by identifying the presence of particular key words (e.g., “sorry”, “i don’t know”) in the system’s response. This rule-based feature does not generalize to a system that uses different set of phrases to indicate its inability to satisfy user’s request. Even temporal dialogue level features computed over turns ($t_1:t_n$) were also hand-crafted and computed by taking simple aggregate statistics (e.g., mean) over turn level features.

3.1 LSTM-based Response Quality Estimation Models

In order to overcome the limitation of hand-crafting temporal features, we propose using a Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) based model to estimate turn-level satisfaction ratings sequentially on a continuous [1–5] scale. Rach et al. (2017) showed that by using only turn-level features, pre-computed temporal features were no longer required for estimating IQ using a LSTM network. To keep the turn-level dialogue quality estimation system causal (Li Tan, 2013), where the output at the current time step only depends on current and previous steps, we do not introduce bi-directionality (Graves et al., 2013) into the network architecture (See Figure 2). Unlike dialogue-level rating, which is computed at the end of a dialogue-session, only past dialogue-context is available to compute a turn’s quality rating. Causality enables using turn-level model to optimize dialogue policies online.

Models encoding sentences into embedding vectors have been successfully used in transfer learning and performing several downstream Natural Language Processing tasks (e.g., Classification and semantic textual similarity detection). Pre-trained sentence representations provided by Universal Sentence Encoder (USE) (Cer et al., 2018) model achieved excellent results on semantic relatedness and textual similarity tasks (Perone et al., 2018).

To address the limitation with using features derived from hand-crafted rules, we use feature sets which are derived from USE pre-trained (512 dimensional) embeddings from its transformer variant. We introduce a set of five features derived

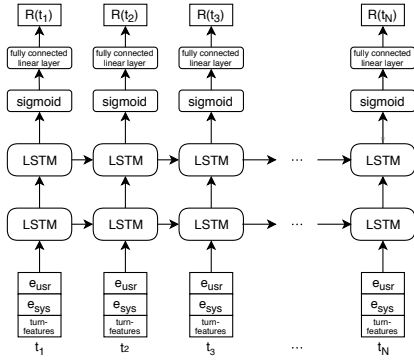


Figure 2: Uni-directional LSTM model to predict RQ ratings at each time-step to estimate dialogue quality at turn-level. e_{usr} , e_{sys} and *turn-features* are pre-trained Universal Sentence Encoder embeddings for user request, system response and rest of the features in Table 2 respectively.

Feature name	Methodology used to compute the feature
ASR Conf. score	Available in the output of the ASR system
NLU Conf. score	Available in the output of the NLU system
Barge-in	Output from ASR
USE embedding of user request	USE embeddings of t_n^u
USE embedding of system response	USE embeddings of t_n^s
NLU intent similarity	Sim. between NLU predicted intents for t_n^u and t_{n+1}^u
Semantic paraphrase of user req.	Cosine sim. between USE embeddings of t_n^u & t_{n+1}^u
Syntactic paraphrase of user req.	Jaccard sim. between words in t_n^u & t_{n+1}^u
Semantic req. & resp. coherence	Cosine sim. between USE embeddings of t_n^u & t_n^s
Syntactic req. & resp. coherence	Jaccard sim. between words in t_n^u & t_n^s
Semantic resp. repetition	Cosine sim. between USE embeddings of t_n^s & t_{n-1}^s
Syntactic resp. repetition	Jaccard sim. between words in t_n^s & t_{n-1}^s
Length of User utterance	Number of words in t_n^u
Length of resp.	Number of words in t_n^s
Duration between utterances	Seconds elapsed between t_n^u & t_{n+1}^u
Domain popularity	Avg. # of reqs. per user for predicted NLU domain t_n^u
Intent popularity	Avg. # of reqs. per user for predicted NLU intent t_n^u

Table 2: Turn level features for turn- t_n and the methodology used to compute them. In bold are features derived from USE embeddings that we introduced. Rest of the turn-level features are similar to Bodigutla et al. (2019a) (Appendix Table 11). Note $\sim 65\%$ relative drop in number of features ($48 \rightarrow 17$). resp., conf., avg., sim., #, & req. indicate response, confidence, average, similarity, count and request respectively.

from USE embeddings of user request and system response texts (See Table 2). These features are then concatenated with turn-level features obtained from the SLU (e.g., ASR confidence score), dialogue manager (e.g., system response) output and predicted intent and domain popularity statistics. Concatenated features are passed as input to each time-step of the uni-directional turn-level satisfaction estimation deep LSTM network (Figure 2), that minimizes mean square error loss between actual and predicted turn-level RQ ratings.

4 Dialogue-level Quality Estimation

In this section we discuss the novel joint turn and dialogue quality estimation approach.

4.1 Joint Estimation of Turn and Dialogue Level Ratings

Turn-level satisfaction estimation helps identify a particular turn’s success from an end user’s perspective. In addition to predicting whether individual turn was successful, we need a dialogue level user satisfaction metric for learning dialogue policies that maximize end user satisfaction on the overall dialogue. Dialogue-level metric also helps in identifying problematic dialogues which caused dissatisfaction to the end user.

We propose a novel approach (Figure 3) to jointly predict turn and dialogue level satisfaction ratings for a given dialogue. Unlike turn-level satisfaction estimation, we are not constrained to use only historical context of a dialogue to predict the dialogue-level ratings as entire context of the dialogue is available while predicting a dialogue level rating. Hence instead of LSTMs we use deep BiLSTM (Graves et al., 2013) network for the dialogue-level satisfaction estimation task. Ultes (2019) showed that BiLSTMs with self-attention (Zheng et al., 2018) model gave the best performance on the IQ prediction task and the model implicitly encoded temporal dependencies. Feature inputs to the joint model are same as the ones we use for turn-level quality estimation in Section 3.1.

Individual turn’s predicted RQ rating does not provide enough information to estimate whether an entire dialogue is satisfactory. Bodigutla et al. (2019b) used average turn-level predicted RQ ratings as feature to estimate dialogue-level quality. We hypothesize that users do not equally weigh each each turn’s success (or failure) while determining end dialogue rating (Example conversation in Appendix Table 8). We apply attention (Vaswani et al., 2017) over turn-level ratings and concatenate the aggregate weighted turn-level rating with the entire dialogue’s representation (hidden state h_{t_N} in Figure 3) before passing it through the sigmoid activation layer for dialogue rating prediction.

In the next section we describe the multi-task loss function we minimized for jointly estimating turn and dialogue-level quality ratings.

4.2 Multi-task Loss Function for Joint Turn and Dialogue Quality Estimation

RQ ratings provided by experts are reliable and consistent (Bodigutla et al., 2019a), however user ratings at the end of a dialogue in general are noisy and it is not clear if they would be coopera-

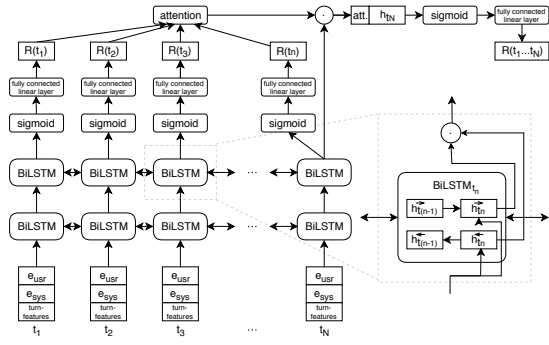


Figure 3: BiLSTM based joint turn and dialogue-level satisfaction estimation model.

tive enough to provide correct feedback (Su et al., 2015). To address the difference in noisiness of labels provided for each task, we followed the approach by Kendall et al. (2017) to use homoscedastic (task-dependent) uncertainty to weigh losses from two tasks, where multi-task loss function is derived by maximizing Gaussian likelihood with homoscedastic uncertainty (Equation 1). Sufficient statistics $f^W(x)$ is the output of a neural network with weight W on input x . Y_t (turn-ratings) and Y_d (dialogue-ratings) are model outputs.

$$p(y_t, y_d | f^W(x)) = p(y_t | f^W(x)) \cdot p(y_d | f^W(x)) \\ = \mathcal{N}(y_t; f^W(x), \sigma_t^2) \cdot \mathcal{N}(y_d; f^W(x), \sigma_d^2) \quad (1)$$

$$\mathcal{L}(w) = -\log p(y_t, y_d | f^W(x)) \\ \propto \frac{1}{2\sigma_t^2} \|y_t - f^W(x)\|^2 + \|y_d - f^W(x)\|^2 + \log \sigma_t + \log \sigma_d \\ = \frac{1}{2\sigma_t^2} \mathcal{L}_t(W) + \frac{1}{2\sigma_d^2} \mathcal{L}_d(W) + \log \sigma_t + \log \sigma_d \quad (2)$$

Equation 2 shows the multi-task loss function \mathcal{L} we minimize. \mathcal{L}_t and \mathcal{L}_d are the mean square error losses computed on turn-level RQ ratings and dialogue-level user ratings respectively. Minimizing the objective functions with respect to noise parameters σ_t and σ_d is interpreted as learning the weights for \mathcal{L}_t and \mathcal{L}_d adaptively from the data. Higher the noise, lower is the weight of the corresponding loss. This method to weigh the losses using learnt weights helps in bringing the losses from the two tasks on the same scale as well.

5 Data and Experimental Setup

This section describes our turn and dialogue-level datasets and explains our experimentation setup.

5.1 Dialogue Quality Data

In order to test the generalizability of the turn and dialogue level user satisfaction models across different domains, user groups and dialogue systems, we sampled 3,129 dialogue sessions (20,167 turns)

Dialogue System	# Domains	# Dialogues	# Turns	Avg. # Turns per Dialogue
A	24	2,133	10,774	5
B	4	996	9,393	9.5

Table 3: Stats on dialogues from dialogue systems A & B from 28 domains (Table 3). These multi-domain dialogues (Example goals user try to achieve in Appendix Table 10) are representative of end user interactions with Alexa and were randomly sampled from two dialogue systems.

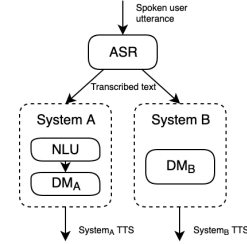


Figure 4: Dialogue systems A & B with their own dialogue managers to process user request and generate Text To Speech (TTS) response once the shared ASR component does the speech to text translation.

Dialogue-system A uses a pipelined modular dialogue agent comprising of ASR, NLU, State Tracker, Dialogue Policy and Natural Language Generation components (Williams et al., 2016). Dialogue-system B is an end-to-end neural model (Ritter et al., 2011; Shah et al., 2018) that shares only the ASR component with system A (Fig. 4).

Each turn was rated by expert RQ annotators² and Dialogue level ratings were provided by end users. Users provided their satisfaction rating with the dialogue on a discrete [1 – 5] scale at the end of each session, irrespective of the outcome. Similar to Bodigutla et al. (2019b) the rating scale we asked the users to follow was 1=Very dissatisfied, 2=Dissatisfied, 3=Moderately Satisfied (or Slightly dissatisfied), 4=Satisfied and 5=Extremely Satisfied. Since earlier attempts to estimate explicit dialogue-level satisfaction ratings did not generalize to different user population (see section 1), we collected dialogue ratings from users belonging to “novice” (15%), “some experience” (33%) and “experienced” (52%) groups. A novice user has minimal experience conversing with the SDS and he/she has never used the functionality provided by the 28 domains prior to the study. A user with some experience has interacted with some (but not all) domains, whereas an experienced user is a seasoned user of Alexa and its domains.

² Expert RQ annotators consistently achieve a high agreement (correlation ≥ 0.8) with other expert annotators and with explicit turn-level user ratings collected through user studies.

5.2 Experimental Setup

This section describes the experimental setup we used for training and evaluating turn and dialogue level satisfaction estimation models.

5.2.1 Turn-level Dialogue Quality Estimation

Similar to Bodigutla et al. (2019a), we considered regression models for experimentation to predict turn-level satisfaction rating on a continuous [1–5] scale. We experimented with two variants of the turn-level satisfaction estimation model described in Section 3.1. In the first variant ($LSTM_{embedding}$) we passed concatenated pre-trained USE sentence embeddings of the user request and system response as input to each time step of the LSTM based model. In the second variant ($LSTM_{embeddings \odot features}$) we concatenate USE embeddings with rest of the 15 turn-level features mentioned in Table 2. We benchmarked the performance of the two LSTM models against the best performing (Bodigutla et al., 2019a) turn-level Gradient Boosting Regression model trained with 48 hand-crafted features (Appendix Table 11).

5.2.2 Dialogue-level Quality Estimation

We experimented with eight models to estimate dialogue level user satisfaction ratings. Three out of the eight models were used as baseline models, which are: 1) Gradient Boosting Regression ($G.Boost$) model trained using features derived from the entire dialogue context ($t_{1:N}$), including hand-crafted turn-level and temporal features (See Appendix Table 11); 2) Two-layer BiLSTM model ($BiLSTM_{features}$) trained with all turn-level features (Table 2), except for the embeddings themselves; 3) $BiLSTM_{features}$ model with self-attention mechanism ($BiLSTM_{features}^{attn}$), which is also a variant of best performing IQ estimation model (Ultes, 2019). For benchmarking we used best performing (Bodigutla et al., 2019b) $G.Boost_{RQ}$ dialogue-level quality estimation model, which used average predicted RQ rating as an additional feature to train the $G.Boost$ model.

Remaining four models we experimented with comprised of two variants of our proposed BiLSTM based joint dialogue quality estimation model, that used attention over the predicted RQ ratings to predict dialogue level rating (See Section 4.1). First variant used only USE embeddings as features ($Joint_{embeddings}^{attn}$) and the second one ($Joint_{embeddings \odot features}^{attn}$)

used all the turn-level features mentioned in Table 2. To test whether including USE embeddings on user request and system response texts alone improved the performance of the baseline $BiLSTM_{features}$ and $BiLSTM_{features}^{attn}$ models, we experimented with their respective counterparts $BiLSTM_{embeddings \odot features}$ and $BiLSTM_{embeddings \odot features}^{attn}$ models that included USE embeddings as features.

The joint models minimized adaptive weighted loss (Eq. 2). All the deep neural models we experimented with used Adam (Kingma and Ba, 2014) optimizer with learning rate 0.0001, mini-batch size of 64 and hidden vector size 512. We used early stopping criteria and (0.5) dropout (Srivastava et al., 2014) regularization techniques to avoid overfitting. Hyper-parameter ranges we experimented with are in Appendix Table 12.

For both dialogue and turn level quality estimation, dialogues were randomly split into training (80%), validation (10%) and test (10%) sets, so that turns from the same dialogue do not appear in both test and training sets.

We trained and evaluated the performance of the turn and dialogue-level quality estimation models on dialogues from dialogue-system A and from both systems A & B combined³. In the first case we used all turn-level features mentioned in Table 2. In the second case we excluded features derived from NLU as dialogue-system B did not use NLU output.

5.2.3 Evaluation Criteria

We used Pearson’s linear correlation coefficient (r) for evaluating each model’s 1-5 prediction performance. For the use case to identify problematic turns from an end user’s perspective, it is sufficient to identify *satisfactory* (rating ≥ 3) and *dissatisfactory* (rating < 3) interactions (Bodigutla et al., 2019b). We used F-score for the dissatisfactory class as the binary classification metric, as most turns and dialogues belong to the satisfactory class. Dialogue-level ratings have a smoother distribution (Pearson’s moment coefficient of skewness -0.27) over turn-level RQ ratings (skewness -0.64).

³ For completeness we evaluated dialogue-quality estimation results using train and test dialogues from System B (Appendix Table 13). Due to limited data (96 test dialogues) performance comparison between models is inconclusive and needs further experimentation.

6 Results and Analysis

This section presents the turn and dialogue-level user satisfaction estimation results.

Model\Metric	System A		Systems A & B	
	Correlation	F - dissat	Correlation	F - dissat.
Gradient Boosting Regression	0.77 ± 0.02	0.77 ± 0.02	0.74 ± 0.02	0.72 ± 0.02
$LSTM_{embedding}$	0.78 ± 0.02	0.80 ± 0.03	0.74 ± 0.03	0.77 ± 0.03
$LSTM_{embedding \odot features}$	0.79 ± 0.03	0.81 ± 0.02*	0.76 ± 0.02	0.78 ± 0.02*

Table 4: Turn-level dialogue quality estimation models’ performance measured using correlation between predicted and true ratings and F-score on dissatisfactory class (F-dissat.) Cells show the mean and 95% bootstrap confidence interval, highest mean in bold, * for statistically significant improvement over benchmark Gradient Boosting Regression model’s performance.

6.1 Turn-level satisfaction Estimation

As shown in Table 4, our proposed LSTM based turn-level quality estimation model outperformed the benchmark Gradient Boosting regression model and removed the need to hand-craft features. Even when NLU features were not used, on dialogues from both dialogue systems, the best-performing ($LSTM_{embedding \odot features}$) model achieved $\sim 3\%$ relative improvement in correlation ($0.74 \rightarrow 0.76$) and statistically significant (at 95% bootstrap-confidence interval) relative improvement 8.3% ($0.72 \rightarrow 0.78$) in F-score on dissatisfactory class performance, over the benchmark model.

6.1.1 Analysis of turn-level model’s performance on new domain

To further test the generalizability of the $LSTM_{embedding \odot features}$ model to new domains, we wanted to verify that the model was not overfitting domain specific vocabulary. To achieve this, we trained the turn-level model with varying percentage of dialogues from a new “movie reservation & recommendation” domain hosted on dialogue System A. Training set consisted of dialogues from Systems A&B and specified percentage of dialogues from the new domain⁴. Consistent with the results in (Bodigutla et al., 2019b), the prediction performance dropped when no dialogues from the new domain were in the training set (results in Appendix Table 14). However, when the model was trained with (randomly sampled) mere 10% (9% train, 1% validation) of dialogues ($\sim 6\%$ slot-value coverage⁵), the prediction performance on *F - dissatisfactory* metric (0.75 ± 0.01) was at

⁴ Since System B did not use NLU, it is not possible to train the model with utterances de-lexicalized using NLU output, such as predicted Intents and slots (Tur and De Mori, 2011).

⁵ Slot-value coverage is the % of unique (slot-type, value) pairs for the specific domain in the selected set of dialogues.

par (difference not statistically significant) with the overall performance achieved by the model when it was trained with 90% (80% train, 10% valid) dialogues (Table 4). Performance parity in-terms of *Correlation* (0.74 ± 0.03) was achieved when $LSTM_{embedding \odot features}$ model was trained with 60% (54% train 6% validation) of dialogues ($\sim 50\%$ slot-value coverage). These two observations imply that binary prediction performance improvement requires training with fewer dialogues in comparison to the number of dialogues required to accurately identify the degree of user (dis)satisfaction.

In order to further understand the relationship between slot types and annotated RQ labels we calculated the Pointwise Mutual Information (PMI⁶) score for the new domain, between its 8 slot-types and 5 RQ labels (total 40 values). Most of the dissatisfactory turns were associated with the system not interpreting the theater names (slot-types ‘theater’) and instructions containing numbers (e.g., “pick the fourth one”) correctly. Validating our hypothesis that users do not perceive all turns’ failures equally, based on the PMI scores, users seem more dissatisfied with system’s failure to identify “theater” (RQ rating - 1) over failure in interpreting numeric instructions (RQ rating - 2)⁷. We calculated cosine similarity between the 40 dimensional PMI scores vector of ($Slot_{type}, RQ_{labels}$) in each selected training set, with PMI scores vector computed on entire set of dialogues in the new domain. As shown in Appendix Table 14, the turn-level model’s performance on new domain improves with the similarity score. This observation suggests that the model is not overfitting to domain specific vocabulary (e.g., movie name), instead it learns the extent of user (dis)satisfaction to failures/success of different (slot) types of requests he/she makes.

6.2 Dialogue-level Satisfaction Estimation

As shown in Table 5, on test sets from System A and System A & B combined, $Joint_{embeddings \odot features}^{attn}$ model outperformed the seven other models we experimented with. On test dialogues from System A & B, in comparison to the baseline $BiLSTM_{features}^{attn}$ model, the Joint-model achieved statistically significant (at 95% confidence interval) absolute 27% ($0.43 \rightarrow 0.70$) improvement in correlation and 17% (0.51

⁶ PMI of pair of outcomes (x, y) belonging to discrete random variables X, Y is $\log \frac{p(x,y)}{p(x)p(y)}$.

⁷ Since RQ ratings are highly correlated with turn-level user satisfaction ratings (Bodigutla et al., 2019a).

→ 0.68) in F-score on dissatisfactory class. In comparison to benchmark $G.Boost_{RQ}$ model, the absolute improvement on the same metrics was 7% and 5% respectively. Learnt noise ratio of 1.2 between the two learnt parameters σ_d^2 and σ_t^2 (Eq. 2), shows higher variance in dialogue-level ratings over turn-level labels.

Model/Metric	System A		Systems A & B	
	Correlation	F - dissatisfactory	Correlation	F - dissatisfactory
$G.Boost$	0.59 ± 0.06	0.61 ± 0.08	0.60 ± 0.05†	0.63 ± 0.04*†
$G.Boost_{RQ}$	0.66 ± 0.05†	0.66 ± 0.06	0.63 ± 0.06*†	0.63 ± 0.05*†
$BiLSTM_{features}$	0.54 ± 0.07	0.63 ± 0.08	0.48 ± 0.08	0.51 ± 0.06
$BiLSTM_{embeddings \odot features}$	0.62 ± 0.07†	0.60 ± 0.07	0.66 ± 0.06*†	0.66 ± 0.05*†
$BiLSTM_{attn_{features}}$	0.44 ± 0.10	0.51 ± 0.09	0.43 ± 0.08	0.51 ± 0.06
$BiLSTM_{attn_{embeddings \odot features}}$	0.61 ± 0.08	0.64 ± 0.06	0.59 ± 0.06†	0.67 ± 0.04*†
$Joint_{attn_{embeddings}}$	0.68 ± 0.08†	0.65 ± 0.07	0.68 ± 0.06*†	0.67 ± 0.06*†
$Joint_{attn_{embeddings \odot features}}$	0.69 ± 0.07*†	0.71 ± 0.07†	0.70 ± 0.06*†	0.68 ± 0.05*†

Table 5: Performance of dialogue-level quality estimation models⁸. Each cell shows the mean and 95% bootstrap confidence interval with the highest mean in bold. * and † indicate statistically significant performance in comparison to baseline $BiLSTM_{features}$ and $BiLSTM_{attn_{features}}$ models respectively. Compared to Table 4, wider confidence intervals are due to sparsity of dialogue-level ratings (~ 15% of turn-level ratings).

Including USE embeddings as features improved the performance of the dialogue-level satisfaction estimation models. Specifically on data from both systems, both $BiLSTM_{embeddings \odot features}$ and $BiLSTM_{attn_{embeddings \odot features}}$ models achieved around absolute 15% - 18% significant improvement in both correlation and F-score on dissatisfactory class performance over their respective counterparts $BiLSTM_{features}$ and $BiLSTM_{attn_{features}}$.

6.2.1 Analysis of learnt Attention Weights

For the $Joint_{attn_{embeddings \odot features}}$ model, Table 6 shows the attention weights learnt on predicted turn level (\hat{RQ}) and true RQ ratings for each turn of a sample dialogue. The joint model puts more weight on the dissatisfactory turns over the satisfactory ones and the dialogue was correctly identified as dissatisfactory. Table 7 shows an example dialogue where a generous (Kulikov et al., 2018) user was satisfied (dialogue-rating 4.0) even when the system did not offer any alternate time slots or restaurant suggestions when his/her initial request to book a table could not be fulfilled. However, the model predicted the dialogue as dissatisfactory.

7 Conclusions

In this paper, we proposed a novel approach to use annotated consistent turn-level Response Quality (RQ) ratings for dialogue level user satisfac-

⁸ Results are not broken down further by domain, since a multi-domain conversation session comprises of turns which belong to \geq one domain and context is shared between them.

⁹ This dialogue is not a real user conversation with live system.

Utterance	Attention weight and turn-level ratings
User: Open {Anonym. movie app.}	System: Hey this is {application} { W_{attn} : 0.073, \hat{RQ} : 5.0, RQ : 5.0}
User: tell me the plot of hustlers	System: Here is a summary of Hustlers ... { W_{attn} : 0.077, \hat{RQ} : 4.3, RQ : 5.0}
User: Give me the whole cast	System: The main role was played by ... { W_{attn} : 0.087 , \hat{RQ} : 1.8, RQ : 2.0}
User: What is it rated?	System: It has a rating of 6.6 on i.m.d.b { W_{attn} : 0.075, \hat{RQ} : 4.7, RQ : 5.0}
User: Exit	System: {Exits the app} { W_{attn} : 0.079, \hat{RQ} : 3.6, RQ : 5.0}
User: play hustler soundtrack	System: Heres a station for you based on The Hustler by Willie ... { W_{attn} : 0.090 , \hat{RQ} : 1.1, RQ : 1.0}
User: Stop	System: {Plays the station} { W_{attn} : 0.090 , \hat{RQ} : 1.1, RQ : 1.0}
User: Open {Anonym. tkt-booking app}	System: Welcome to {application} { W_{attn} : 0.078, \hat{RQ} : 4.0, RQ : 5.0}
User: Where is hustlers playing ?	System: You can go to AMC Which theater would you like? { W_{attn} : 0.087 , \hat{RQ} : 1.6, RQ : 2.0}
User: AMC	System: Today at AMC: Joker... which movie would you like to see? { W_{attn} : 0.088 , \hat{RQ} : 1.6, RQ : 2.0}
User: Is the hustlers playing there today	System: Today at AMC: Joker... which movie would you like to see? { W_{attn} : 0.087 , \hat{RQ} : 1.7, RQ : 2.0}
User: No	System: {No response} { W_{attn} : 0.089 , \hat{RQ} : 1.3, RQ : 2.0}

Table 6: Conversation⁹ where the $Joint_{attn_{embeddings \odot features}}$ correctly predicts the dialogue as dissatisfactory (rating = 2.5). W_{attn} , \hat{RQ} and true RQ ratings indicate learnt attention weights, predicted and true RQ ratings respectively. Weights in bold indicate top 50% of learnt attention weights.

Utterance	Attention weight and turn-level ratings
User: Open {restaurant booking app.}	System: Hey this is {application} { W_{attn} : 0.126, \hat{RQ} : 5.0, RQ : 5.0}
User: Can you can you uh book a table at koreana for two ?	System: Sorry, this is not curr. supported. { W_{attn} : 0.147 , \hat{RQ} : 1.8, RQ : 2.0}
User: Can you book a table at the cheesecake factory for three ?	System: What time would you like to make the reservation for? { W_{attn} : 0.137, \hat{RQ} : 3.2, RQ : 4.0}
User: Three oclock	System: What date would you like the reservation for? { W_{attn} : 0.135, \hat{RQ} : 3.6, RQ : 4.0}
User: For tomorrow	System: Sorry. No restaurants could be found that match your criteria. { W_{attn} : 0.152 , \hat{RQ} : 1.1, RQ : 2.0}
User: Ok	System: {No response} { W_{attn} : 0.153 , \hat{RQ} : 1.0, RQ : 4.0}
User: Stop	System: {No response} { W_{attn} : 0.149 , \hat{RQ} : 1.5, RQ : 4.0}

Table 7: Conversation⁹ where the model incorrectly predicts the dialogue as defective (rating 2.3). User’s rating is 4.0.

tion estimation on conversations which span three user groups, 28 domains and two dialogue systems. With the help of pre-trained Universal Sentence Encoder (USE) embeddings, we removed the need to hand-craft features. Leveraging noise adaptive weighting of multi-task loss technique and aggregating predicted RQ ratings using attention mechanism, we developed the BiLSTM based deep joint turn & dialogue level satisfaction estimation model. The best-performing joint-model achieved up to 27% absolute significant improvement in correlation (Pearson’s- r) performance and 7% absolute improvement over the baseline deep neural network and the benchmark $G.Boost$ models, respectively.

Analysis of the learnt attention weights showed that the joint model exhibited the desired behavior to weigh successful and failed turns unequally. However the model was not calibrated to factor in user’s preferences and biases, which we plan to address in future work.

Acknowledgments

We thank Arindam Mandal and Jean-Jacques Loesch for their guidance and support. We thank Alexa Data Services-RAMP team for data. We also thank EMNLP reviewers for their feedback.

References

- L. E. Asri, H. Khouzaimi, R. Laroche, and O. Pietquin. 2014. [Ordinal regression for interaction quality prediction](#). In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3221–3225.
- Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019b. Multi-domain conversation quality evaluation via user satisfaction estimation. In *3rd Workshop on Conversation AI: Today’s Practice and Tomorrow’s Potential, 33rd Conference on Neural Information Processing Systems*.
- Praveen Kumar Bodigutla, Longshaokan Wang, Kate Ridgeway, Joshua Levy, Swanand Joshi, Alborz Geramifard, and Spyros Matsoukas. 2019a. Domain-independent turn-level dialogue quality evaluation via user satisfaction estimation. In *Special session on Implications of Deep Learning for Dialog Modeling, 20th Annual SIGdial Meeting on Discourse and Dialogue*.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. [Universal sentence encoder](#). *CoRR*, abs/1803.11175.
- Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echevoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. [Survey on evaluation methods for dialogue systems](#). *CoRR*, abs/1905.04071.
- Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. [Modeling user satisfaction with hidden Markov models](#). In *Proceedings of the SIGDIAL 2009 Conference*, pages 170–177, London, UK. Association for Computational Linguistics.
- Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Alex Graves, Navdeep Jaitly, and Abdel rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional lstm. In *In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.
- Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Issues in predicting user satisfaction transitions in dialogues: Individual differences, evaluation criteria, and prediction models. In *IWSDS*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural Comput.*, 9(8):1735–1780.
- Alex Kendall, Yarin Gal, and Roberto Cipolla. 2017. [Multi-task learning using uncertainty to weigh losses for scene geometry and semantics](#). *CoRR*, abs/1705.07115.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Ilya Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a search strategy in neural dialogue modelling. *ArXiv*, abs/1811.00907.
- Jean Jiang Li Tan. 2013. Causal systems. *Digital Signal Processing (Second Edition), 2013*, pages 57–85.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Christian S. Perone, Roberto Silveira, and Thomas S. Paula. 2018. [Evaluation of sentence embeddings in downstream and linguistic probing tasks](#). *CoRR*, abs/1806.06259.
- Louisa Pragst, Stefan Ultes, and Wolfgang Minker. 2017. [Recurrent Neural Network Interaction Quality Estimation](#), pages 381–393. Springer Singapore, Singapore.
- C. Preston and A. Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104:1–15.
- Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2017. Interaction quality estimation using long short-term memories. In *SIGDIAL Conference*.

- Alan Ritter, Colin Cherry, and William B. Dolan. 2011. [Data-driven response generation in social media](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. [Agenda-based user simulation for bootstrapping a POMDP dialogue system](#). In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152, Rochester, New York. Association for Computational Linguistics.
- Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011. Modeling and predicting quality in spoken human-computer interaction. In *SIGDIAL Conference*.
- 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 Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey. European Language Resources Association (ELRA).
- Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. [Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.
- Weiyang Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *ACL*.
- Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958.
- Pei-hao Su, David Vandyke, Milica Gasic, Dongho Kim, Nikola Mrksic, Tsung-Hsien Wen, and Steve J. Young. 2015. [Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems](#). *CoRR*, abs/1508.03386.
- G. Tur and R. De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Wiley.
- Stefan Ultes. 2019. [Improving interaction quality estimation with BiLSTMs and the impact on dialogue policy learning](#). In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 11–20, Stockholm, Sweden. Association for Computational Linguistics.
- 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*, pages 303–312, New York, NY. Springer New York.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *CoRR*, abs/1706.03762.
- Marilyn Walker, Candace Kamm, and Diane Litman. 2000. [Towards developing general models of usability with paradise](#). *Nat. Lang. Eng.*, 6(3-4):363–377.
- Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *D&D*, 7:4–33.
- Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. [Opentag: Open attribute value extraction from product profiles](#). In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’18*, pages 1049–1058, New York, NY, USA. ACM.

A Appendices

Utterance	Turn Rating
User: I would like to book a cab to work	
System: Sure, at what time would you like to be picked up?	5.0
User: Eight a.m tomorrow	
System: Could you confirm your cab booking to {office address} at 8 a.m tomorrow?	
You will be charged \$50 for this ride	5.0
User: No, cancel my ride	
System: Ok, your booking is confirmed.	
Your card ending xxxx has been charged \$50	1.0

Table 8: An example of a failed dialogue where the overall dialogue rating is **1.0**. However, average turn ratings **3.7** indicates a successful dialogue.

Rating	Description
1	Terrible (system fails to understand and fulfill user's request)
2	Bad (understands the request but fails to satisfy it in any way)
3	OK (understands users request and either partially satisfies the request or provides information on how the request can be fulfilled)
4	Good (understands and satisfies the user request, but provides more information than what the user requested or takes extra turns before meeting the request)
5	Excellent (understands and satisfies user request completely and efficiently)

Table 9: RQ rating guidelines

Goal	Domains
Get ratings of movies directed by the director of a movie playing in theaters	Movie recommendations and Reservations
Ask for a general type of recipe and then add the ingredients to the shopping list	Recipe, Shopping
Find out the weather in a location and book a ticket to a movie playing in theaters near by	Weather, Location, Movie Recommendations and Reservations
Playing sound track of a popular artist	Knowledge and Music
Book a cab and add a notification for the same	Notifications and Cab booking
Planning activities for eventing	Weather, Restaurants and Cab booking

Table 10: Example goals users tried to achieve and their corresponding domains.

Index	Feature set description	Turn(s) the feature is computed on
1	ASR Confidence	t_n^u
2	NLU Confidence	t_n^u
3	Barge-in	t_n^u
4	Intent popularity computed on predicted NLU intent	t_n^u
5	Domain popularity computed on predicted NLU intent	t_n^u
6	NLU Intent similarity between consecutive turns	$t_n^u-t_{n+1}^u$
7	Syntactic similarity between consecutive turns user utterances	$t_n^u-t_{n+1}^u$
8	Syntactic similarity between user utterance & system response	$t_n^u-t_n^s$
9	Syntactic similarity between current response & previous turn's system response	$t_{n-1}^s-t_n^s$
10	Affirmation prompt in user request	t_n^u
11	Negation prompt in user request	t_n^u
12	Question prompt in user request	t_n^u
13	Termination prompt in user request	t_n^u
14	Next turn's ASR Confidence	t_{n+1}^u
15	Next turn's NLU Confidence	t_{n+1}^u
16	Next turn's Barge-in indicator	t_{n+1}^u
17	Affirmation prompt in next turn's user request	t_{n+1}^u
18	Negation prompt in next turn's user request	t_{n+1}^u
19	Question prompt in next turn's user request	t_{n+1}^u
20	Termination prompt in next turn's user request	t_{n+1}^u
21	Intent popularity computed on next turn's predicted NLU intent	t_{n+1}^u
22	Domain popularity computed on next turn's predicted NLU intent	t_{n+1}^u
23	Affirmation prompt in system response	t_n^s
24	Negation prompt in system response	t_n^s
25	Question prompt in system response	t_n^s
26	Un-actionable user request	t_n^s
27	# Un-actionable user request	$t_1^s-t_n^s$
28	# Barge-ins	$t_1^u-t_n^u$
29	# Question prompt in system response	t_n^s
30	# Negation prompt in system response	t_n^s
31	# Affirmation prompt in system response	t_n^s
32	# Termination prompt in user request	t_n^u
33	# Question prompt in user request	t_n^u
34	# Negation prompt in user request	t_n^s
35	# Unique Intents/# Length of dialogue so far	t_1-t_n
36	Length of the dialogue so far	t_1-t_n
37	Avg ASR confidence	$t_1^u-t_n^u$
38	Avg NLU confidence	$t_1^u-t_n^u$
39	Avg Semantic similarity between consecutive turns' user utterances	$t_1^u-t_{n+1}^u$
40	Avg Syntactic similarity between consecutive turns' user utterances	$t_1^u-t_{n+1}^u$
41	Avg Semantic similarity between consecutive turns' system responses	$t_1^s-t_n^s$
42	Avg Syntactic similarity between consecutive turns' system responses	$t_1^s-t_n^s$
43	Avg Semantic similarity between user utterance and system responses	$t_1^u-t_n^s$
44	Avg Syntactic similarity between user utterance and system responses	$t_1^u-t_n^s$
45	Avg aggregate - domain popularity	t_0-t_n
46	Avg time difference between consecutive utterances	$t_0^u-t_n^u$
47	Avg aggregate - intent popularity	t_0-t_n
48	Avg aggregate - domain popularity	t_0-t_n

Table 11: Features used for predicting turn ratings by [Bodigutla et al. \(2019a\)](#). # indicates count. Features 10-45 cover dialogue system specific rule based turn-level features and hand-crafted temporal features.

Model	Hyper parameter and their corresponding ranges
Gradient Boosting Decision Trees	max-depth: [2–10], min-samples-leaf: [2–10], min-samples-split: [2–10]
LSTM and BiLSTM Based models for turn and dialogue level quality estimation	n-layers: [1, 2, 3], hidden size: [8, 16, 32, 64, 128, 256, 512, 1024], batch size: [8, 16, 32, 64, 128], optimization: [sgd, Adam, RMSProp], dropout: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7], learning rate: [1.0, 0.1, 0.001, 0.0001, 0.00001], length of the sequence: [9 – 20]

Table 12: Hyper parameter value ranges we used for training turn-level and dialogue-level quality estimation. DNN models were implemented in PyTorch (Paszke et al., 2019) and Gradient Boosting Regression model was implemented using scikit-learn (Preston and Colman, 2000). Parameters were tuned using grid-search and experiments were run on P2.xlarge AWS EC2 compute instance, which has 1 NVIDIA K80 GPU, 4 vCPUs, 61GiB RAM.

Model\Metric	System B	
	Correlation	F – dissatisfactory
<i>G.Boost</i>	0.67 ± 0.10 †	0.64 ± 0.07
<i>G.Boost_{RQ}</i>	0.70 ± 0.09 †	0.64 ± 0.06
<i>BiLSTM_{features}</i>	0.53 ± 0.13	0.66 ± 0.11
<i>BiLSTM_{embeddings}⊙_{features}</i>	0.69 ± 0.09 †	0.66 ± 0.10
<i>BiLSTM_{features}^{attn}</i>	0.38 ± 0.15	0.57 ± 0.14
<i>BiLSTM_{embeddings}⊙_{features}^{attn}</i>	0.68 ± 0.11 †	0.71 ± 0.08
<i>Joint_{embeddings}^{attn}</i>	0.62 ± 0.10	0.72 ± 0.08
<i>Joint_{embeddings}⊙_{features}^{attn}</i>	0.65 ± 0.10 †	0.65 ± 0.09

Table 13: Dialogue-level quality estimation on 96 test dialogues from System B. Dialogues from training (800) and validation (100) were obtained from the same system as well. Models were trained without NLU features. Larger 95% bootstrap confidence intervals around the mean are due to limited test data. † indicates statistical significance over *BiLSTM_{features}^{attn}* model’s results.

%Train Dialogues	F – dissatisfactory	Correlation	Slot-Value coverage	$Cos_sim(PMI_{(slot-type,label)}^{train}, PMI_{(slot-type,label)}^{all-dialogues})$
0	0.68 ± 0.02	0.55 ± 0.03	-	-
10	0.75 ± 0.01	0.67 ± 0.02	6.20%	0.516
20	0.74 ± 0.02	0.68 ± 0.02	13.5%	0.546
30	0.76 ± 0.02	0.69 ± 0.02	21.1%	0.781
40	0.78 ± 0.02	0.73 ± 0.02	31.4%	0.796
50	0.77 ± 0.03	0.72 ± 0.03	40.9%	0.854
60	0.79 ± 0.03	0.74 ± 0.03	48.9%	0.864
70	0.80 ± 0.02	0.74 ± 0.03	58.8%	0.886
80	0.83 ± 0.04	0.77 ± 0.03	70.9%	0.931
90	0.84 ± 0.04	0.78 ± 0.03	83.9%	0.963

Table 14: Turn-level *LSTM_{embeddings}⊙_{features}* model’s performance on multi-domain dialogues consisting of new multi-turn Movie Reservation & Recommendation domain (450 dialogues, 1500 turns). Train dialogues % indicates, the percentage of dialogues (out of 450) used for training (90% train, 10% validation split). Slot-Value coverage is the percentage of unique (slot-type, value) pairs in each training set. *Cos_sim* is the cosine similarity between 40 (8 slot-type x 5 RQ-label categories) dimensional Pointwise Mutual Information (PMI) vectors computed on (slot-type, label) pair from dialogues in training set with PMI vector computed on (slot, label) pair from all 450 dialogues.