IQ-Net: A DNN Model for Estimating Interaction-level Dialogue Quality with Conversational Agents

Yuan Ling, Benjamin Yao, Guneet Kohli, Tuan-Hung Pham, Chenlei Guo
yualing,benjamy,gkohli,hupha,guochenl@amazon.com
Amazon Alexa AI
Seattle, WA

ABSTRACT

An automated metric to evaluate dialogue quality is critical for continuously optimizing large-scale conversational agent systems such as Alexa. Previous approaches for tackling this problem often rely on a limited set of manually designed and/or heuristic features, which cannot be easily scaled to a large number of domains or scenarios. In this paper, we present Interaction-Quality-Network (IQ-Net), a novel DNN model that allows us to predict interaction-level dialogue quality directly from raw dialogue contents and system metadata without human engineered NLP features. The IQ-Net architecture is compatible with several pre-trained neural network embeddings and architectures such as CNN, Elmo, and BERT. Through an ablation study in Alexa, we demonstrate that several variants of IQ-Net outperform a baseline model with manually engineered features (3.89% improvement in F1 score, 3.15% in accuracy, and 6.1% in precision score), while also reduce the efforts to extend to new domains/use-cases.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

online evaluation, intelligent conversational agents (ICAs), evaluation metrics, defect detection

ACM Reference Format:
Yuan Ling, Benjamin Yao, Guneet Kohli, Tuan-Hung Pham, Chenlei Guo. 2020. IQ-Net: A DNN Model for Estimating Interaction-level Dialogue Quality with Conversational Agents. In *Proceedings of KDD Workshop on Conversational Systems Towards Mainstream Adoption (KDD Converse'20).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

As voice-controlled intelligent conversational agents (ICAs), such as Alexa, Siri, and Google Assistant become increasingly popular, ICAs have become a new paradigm for accessing information. They represent a hybrid of search and dialogue systems that conversationally interact with users to execute a wide range of actions (e.g., searching the Web, setting alarms, and making phone calls) [9, 31]. These ICAs are complex systems with many components, such as automatic speech recognition (ASR), natural language understanding (NLU), language generation, and dialog management. As a result, there are generally two categories of evaluation metrics for ICAs [32]: (1) Component metrics which are used to measure the quality of each individual component, such as Word Error Rate (WER) and NLU Accuracy; (2) End-to-End (E2E) metrics that are designed to measure the entire dialogue quality and/or user satisfaction. In this paper, we focus on "online" or "automated" E2E metrics, which are built with machine learning models that are designed to predict user satisfaction. Defining online E2E metrics is a popular subject of various research [10] due to its vital role in continuous optimization of ICA systems [15].

1. Assessment: non-defect (success)
User request: play five little ducks
ICA response: Ok, playing five little ducks
User request: (after 2 mins...) play five little ducks
2. Assessment: defect, user paraphrase
User request: play three little ducks
ICA response: Sorry, I cannot find the song
User request: play five little ducks
3. Assessment: non-defect, user confirmation
User request: turn off light
ICA response: did you mean Lamp One?
User request: yeah
4. Assessment: defect, user correction and ASR error
User request: buy
ICA response: you have one item in your shopping cart {name of the item}, do you want to buy it?
User request: bye-bye.

Table 1: Example dialogues and interaction-quality assessment. Note that assessments are for the 1st turn in each dialogue.

Prior attempts to model online E2E metrics could be roughly grouped into two categories: (1) Dialogue-level metrics such as the popular PARADISE framework [39] which aims to predict dialogue-level satisfaction ratings provided by surveyed users; References [8, 36] formulate dialogue interaction as a reinforcement learning task, which aims to predict reward from dialogue history using different variants of DNN models. (2) Turn-level (or exchange-level) metrics such as Interaction Quality (IQ) [33] which predict per turn dialogue quality either provided by users [13] or annotated by human raters [34]. In particular, IQ models have gained popularity recently because of publicly available benchmarks such as the CMU Go-bus information system [34]. Various methods exist to predict Interaction Quality, for example using Hidden Markov

KDD Converse'20, August 2020,

Models [38], Support Vector Machines [12, 33] and Recurrent Neural Networks [28, 29]. However, these approaches rely heavily on dialog system internal features such as ASR confidence and SLU Semantic Parse. While these features a very effective in a small-scale closed-loop system, they are very unreliable in a large organization like Alexa where many teams constantly updating various components in parallel. For example, ASR-confidence score could have significant shift between two ASR model versions hence will be an unreliable input for E2E online metric, which is, in part, designed to measure ASR's impact to user satisfaction. Therefore, instead of relying on system internal signals, we draw on the intuition that human raters could reliably judge the quality of a turn by looking at the context of the dialogue [4] (without ever needing to know what is the ASR confidence score). Table 1 shows four example dialogues with interaction-quality assessment by human.

While these examples demonstrate how human could easily judge the quality of a turn using dialogue context, they are also non-trivial cases for an ML model to predict. For example, example #1 and #2 share similar user paraphrase structure. But example #1 is non-defective (successful) because ICA response is relevant while #2 is clearly defective because the ICA responded with "sorry, I cannot ...". On the other hand, while example #3 and #4 share similar query/response structure (ICA asking for confirmation), #3 is non-defective because user have a positive confirmation next turn while #4 is a defect because user correction next turn. To capture such a diverse of dialogue patterns, it is clear that we need to leverage semantic meaning of the dialogue context. While it is possible to manually extract features such as "paraphrasing", "cohesion between response and request" as proposed by Bodigutla et al. [4], the complexity of open domain system like Alexa limits the efficacy for such approach. Therefore, in this paper we present Interaction-Quality-Network (IQ-Net), an E2E DNN model that allows us to predict interaction-level dialogue quality directly from raw dialogue contents and system metadata without human engineered NLP features. In contrast to existing related work, we contribute to the IQ modeling literature from the following perspectives.

- Instead of focusing on a few specific tasks and system internal features [21, 33], IQ-Net is a generic interaction quality model that could be used for evaluating Interaction Quality across multiple domains and various systems;
- Unlike previous multi-domain user satisfaction evaluation model [4], which relies on manually engineered dialogue features that are hard to scale, IQ-net is capable of capturing a variety of dialogue patterns and could be easily extend to new domain/use-cases as long as we have annotated examples.

The rest of the paper is organized as follows. Section 2 reviews existing work. Section 3 presents our methods to estimate interaction-level dialogue quality. Section 4 presents our experimental results. We conclude our paper in Section 5.

2 RELATED WORK

In this section, we summarize the related work on evaluation methods/metrics for the search systems and error analysis for the ICAs to put our contributions in context.

Evaluation is a central component for information search systems [19]. For text-based information retrieval, the relevant documents/pages are annotated manually to evaluate the search system performance. Query-based metrics such as the mean average precision (MAP) and normalized discounted cumulative gain (nDCG) [20] are frequently used to evaluate the system performance. However, the human annotation process is expensive and error-prone; in addition, the user's individual intent is commonly not taken into consideration. To alleviate this issue, some research models user satisfaction/behaviors to improve the evaluation of system's performance [1] by incorporating the following signals: 1) user behaviors including clicks, dwell time, mouse movements, scrolling behaviors, and abandonment [16]; 2) context-specific features such as viewport metrics [24], touch-related features, and acoustic signals [23], 3) query-based features, such as query refinement, query length, and frequency in logs. While the metrics for evaluating traditional search system might not be used directly to evaluate ICAs, some of the metric components, such as query refinement, can be adapted to evaluate ICAs.

Compared with text-based information retrieval, voice-based information retrieval is quite different [21, 23] because of two reasons. First, voice-based interactions are conversational; in some scenarios, the user expects that the search system is able to refer to the previous interactions to understand the current request. Second, the voice input could provide automatic speech recognition (ASR) errors to downstream applications and affect user satisfaction negatively. Research on Spoken Dialogue System (SDS) attempts to model user satisfaction at turn level as a continuous process over time [13, 18]. An annotated and standardized corpus, such as Let's Go Bus Information System dataset from CMU [34], was developed for classification and evaluation tasks regarding task success prediction, dialogue quality estimation, and emotion recognition. Based on the dataset, an evaluation metric as Interaction Quality [10, 34] is developed with features related to ASR, Spoken Language Understanding (SLU), and Dialog Manager at exchange, dialog, and window level.

ICAs differ from traditional SDS in that they support personalization and a wide range of tasks. Dialogue systems can be categorized into three groups: task-oriented systems, conversational agents, and interactive question answering systems [10]. ICAs are designed to be able to handle all of these tasks; thus, it makes the evaluation of ICAs very challenging. In addition, as the voice-only ICAs tend to evolve to become the voice-enabled multi-modal ICAs, it become even more complex for evaluation. A recent user study on ICAs attempts to compare the differences regarding features [25], performance, ASR error [17], and user experiences [3] across different ICAs. Surveys with questionnaires [2, 5] are conducted to understand functional and topical use of ICAs by individuals; however, these studies are limited to predefined scenarios of interactions.

Currently, there is limited research on building automatic metrics for evaluating ICAs' performances. Jiang et al. [21] built separate models for evaluating user satisfaction on five domains, including Chat, Device Control, Communication, Location, Calendar, and Weather. The models consider several types of features, including user-system interactions, click features, request features, response features, and acoustic features. This work automated the online evaluation for ICAs. However, the work did not consider the variability of interface and interaction; in addition, its scope is limited to ICAs

on mobile devices and several specific scenarios/domains. Bodigutla et al. [4] introduces a Response Quality annotation schema, which showed high correlation with explicit turn level user satisfaction ratings. This paper developed a method for evaluating user satisfaction at turn level in multi-domain conversations users for ICA using five features: user request rephrasing, cohesion between response and request, aggregate topic popularity, unactionable user request, and the diversity of topics in a session. The turn-level user satisfaction rating is further used as feature to improve dialogue-level satisfaction estimation. Other current research on evaluation of ICAs more focuses on user satisfaction estimation and goal success prediction, which are more suitable for dialog-level or task-level evaluation due to the dialogue style of interactions [15, 22, 30]. A user's frustration in the middle of a task or a dialogue might not be captured. The approach also often lacks interpretability in term of the root causes of user frustration. Finally, it is not obvious how one should define task and session boundaries for ICAs [30]; thus, it is critical to evaluate ICAs at turn level.

The complexity of ICAs' components makes it difficult to determine which component causes an error or user frustration. Researchers have studied system errors in search and dialogue systems. For ICAs, the error root causes can be categorized into groups [31, 32], including ASR errors, NLU errors, unsupported system actions, no language generation, back-end failures, endpoint errors, and uninterpretable inputs. These errors can be the root causes of user reformulating their queries [27, 31].

3 METHODOLOGY

In this section, we present IQ-Net: a DNN model for estimating interaction-level dialogue quality. First, we introduce the overall architecture and training procedure of IQ-Net. Then, we explain how we represent the dialogue context in details. Next, we introduce the system metadata used in the IQ-Net.

3.1 IQ-Net

The IQ-Net model is presented in Figure 1. IQ-Net includes two major components: (1) dialogue context representations and (2) a list of features derived from system metadata.

For modeling dialogue context, we consider user's request text plus response text of ICA in the consecutive turns. As showed in Figure 1, the dialogue context representation part takes current turn request and response (U_1 and R_1), and more requests from following turns ($U_2, U_3, ...$) as inputs. For simplicity, we only consider the next one turn request; thus, the inputs can be represented as $< U_1, R_1, U_2 >$. We will support more following turns in future work. We assume $< U_1, R_1 >$ captures the relevancy between user request and Alexa response and $< U_1, U_2 >$ captures patterns from user's repeat/dialog behavior.

We map the word indicies of U_1, R_1, and U_2 into a fixed dimension of vectors through pre-trained word embeddings [11, 26]. The word embedding representation of U_1, R_1, and U_2 all go through sentence encoder E, which can be pre-trained by individual datasets. We use both CNN encoder E_{CNN} and BERT encoder E_{BERT} as sentence encoders in our experiments for comparisons. We concatenate the hidden representations of h_{U_1} and h_{R_1}, h_{U_1} and h_{U_2} accordingly. The

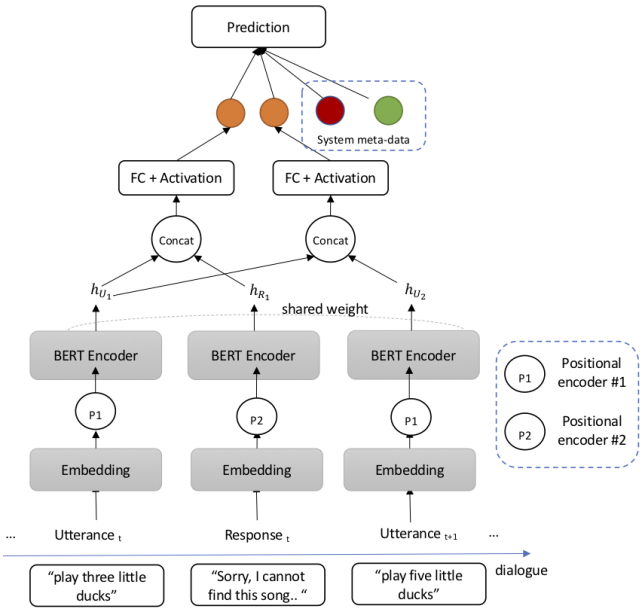


Figure 1: IQ-Net Architecture Diagram. The model considers the information of current turn utterance (U_t), response (R_t), next turn utterance (U_{t+1}). Semantic encoding layers share weights. Position encoders are shared among (U_t) and (U_{t+1}). Note that in this diagram, we use BERT encoder as an example. For IQ-Net(CNN), we replace the BERT encoder with an CNN encoder.

concatenate results for each part followed by a feed-forward network and activation function.

The final outputs from the dialogue context representations will combine with all other features derived from system metadata to predict a defect/non-defect outcome for the interaction-level dialogue quality of the first turn:

$$p(\texttt{Defect} = true | < U_t, R_t, U_{t+1} >, f_M) \quad (1)$$

f_M represents a list of meta-data features (described in Section 3.3).

The objective function for the overall task is

$$L_\Theta = \sum_{(U_t, R_t, U_{t+1})} l(F(U_t, R_t, U_{t+1}, f_M), y) \quad (2)$$

whereas $F()$ is a function that represents IQ-Net. l is the standard cross entropy loss. y is the ground-truth label.

3.2 Dialogue Context Representations

Here, we explain in details that the dialogue context are represented with $< U_1, U_2 >$ modeling and $< U_1, R_1 >$ modeling.

3.2.1 $< U_1, U_2 >$ *Modeling.* $< U_1, U_2 >$ pair contains user's dialog behavior/patterns. For example, ICA users tend to re-express the same intention with follow-up requests after an unsuccessful attempt from the previous request. We refer to the follow-up request as a "rephrase" of the previous request. Identifying rephrasing pattern between requests pair can help discovering defect/frictions.

In addition to the rephrasing behavior between two consecutive user requests, the user can also express the confirm or deny intention in a follow-up request, as shown in Table 2.

Example 1: confirm
User request: Alexa, add paper to my cart
Alexa response: Do you mean paper towel?
User request: Yes.
Example 2: deny
User request: Alexa, add paper to my cart
Alexa response: Do you mean paper towel?
User request: No. add A4 paper to my cart.

Table 2: $< U_1, U_2 >$ confirm/deny examples.

Such patterns existing in $< U_1, U_2 >$ reflect user's real intention through the corresponding repeat/confirm/deny behaviors, which can be learned in the proposed IQ-Net.

3.2.2 $< U_1, R_1 >$ *Modeling.* The semantic relevance of ICA's response and user's request can be an effective feature for defect predictions. When an ICA responds to a user's request with an irrelevant answer, the metric should capture this as defective. However, it is difficult to discover such a defect when the ICA provides a complete but incorrect response, and user chooses to abandon the interaction without rephrasing the request. The relevance between the request and the response text can potentially help with defect identification.

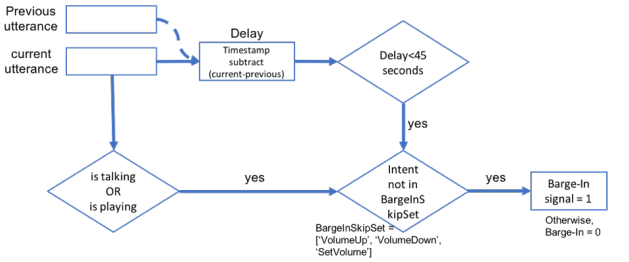
The IQ-Net takes user request and response text (U_1 and R_1) as inputs, similar to $< U_1, U_2 >$ modeling. We adopt the frequently used "Siamese" architecture [6] to measure request-response similarities in the projected space as showed in the Figure 1.

3.3 System Metadata

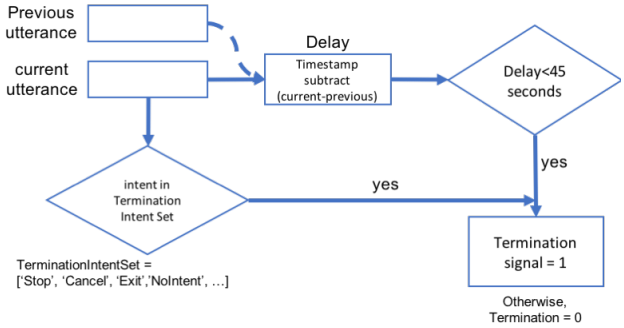
Table 3 introduces our signals and features of system's metadata for evaluating ICAs. We adopt several signals, such as termination and barge-in, to reflect the user's implicit feedback/action. Then, we introduce generic features, such as user domain/intent information from NLU outputs, and some system action types.

Feature	Symbol	Feature Type
[User interruption]		
barge-in	f_{BI}	$\{0, 1\}$ binary values
termination	f_{TM}	$\{0, 1\}$ binary values
gap time	f_{GT}	continuous values
[User intent]		
intent	f_{UI}	categorical feature
domain	f_{UD}	categorical feature
[System Action]		
dialog status	f_{DS}	categorical feature
promptID	f_{PI}	$\{0, 1\}$ binary values
SLU score bin	f_{SB}	categorical feature

Table 3: Feature List Derived from System Metadata.



(a) barge-in



(b) termination

Figure 2: User interruption signals

3.3.1 User Interruption Signals. **User Barge-in:** Barge-in is a frequently used feature for evaluation in SDS [4, 10, 34]. When a customer interrupts a follow-up request while ICA is responding or playing, the turn will be labeled as a barge-in. As shown in Figure 2(a), we build a rule-based barge-in model. When (1) ICA is talking or playing, (2) the delay between the previous utterance and the current one is less than a certain period of time (e.g., 45 seconds), and (3) the user intent is not in the intents set {"VolumeUp","VolumeDown","SetVolume"}, we label the current turn barge-in value $f_{BI} = 1$; otherwise, $f_{BI} = 0$.

User Termination: We define a user termination as when a customer expresses a terminating intent. As shown in Figure 2(b), our termination detection is rule-based. If the user's intent is a terminating action (e.g., StopIntent, ExitAppIntent), the delay between the previous utterance, and the current one is less than a certain period of time (e.g., 45 seconds), we have the termination value $f_{TM} = 1$; otherwise, $f_{TM} = 0$.

Gap Time: The gap time between two requests is an important indicator for a user interruption. We use the time differences as a feature, which is represented as f_{GT}.

3.3.2 User Intent Signals. An NLU component allows ICAs to produce interpretations for an input sentence. The NLU component accepts recognized speech inputs and produces intents, domains, and slots for the input utterance to support the user request [7, 37]. We use the domain and intent outputs from NLU as signals to reflect the user's intention. We cover over dozens of domains and thousands of intents, and use them as categorical features. We use f_{UD} as domain features and f_{UI} as intent features.

3.3.3 System Action Signals. **Dialog Status**: Dialog Management (DM) is a key component of spoken language interactions with ICAs. It makes user inputs actionable by asking appropriate questions to help customers achieve a goal. DM can detect when a valid task completes or if there is trouble in the dialog and it records this information. Following previous work on dialog acts modeling [35], we use DM status values as system action signals for defect detection. Compared with the work [21], we focus on more generic DM status categories here:

- SUCCESS: The ICA is able to act and deliver what it thinks the user wants (not ground truth).
- IN_PROGRESS: The ICA is in the process of executing on a task or is prompting for additional information.
- USER_ABANDONED: The user abandons an in-progress dialog, either explicitly or implicitly.
- INVALID: The Spoken Language Understanding (SLU) could interpret the utterance, but the ICA cannot handle it. For example, the input may express a task that is unactionable due to user dependencies (e.g. account linking for music purchases), or is currently unsupported.
- ICA_ABANDONED: The SLU stops trying / ICA reaches the MAX number of turns.
- FAULT: The ASR encounters some internal errors, the NLU service fails, or the app fails.

We represent the DM status as categorical feature f_{DS}.

System Prompts: promptID is a free form system status code provided by ICA speechlets to indicate whether a speechlet can handle the request. For example, when the ICA responds "Sorry, I'm not sure", the promptID is "NotUnderstood". promptIDs can be categorized and mapped to different types of frictions such as SLU frictions, errors or retries, coverage gaps, unsupported use cases and user actions required. We convert the promptID into a binary feature; if the promptID is mapped to any friction type, the feature value $f_{PI} = 1$, otherwise $f_{PI} = 0$.

SLU Score Bin: The SLU score represents the confidence of what the SLU understoods as the desired intent/slot output for the utterance. The SLU score bin is a categorical feature to group the confidence score into high/medium/low bins. Comparing to volatile features such as "ASR confidence" or "Entity resolution score", SLU score bin a stable feature that has low variation over-time. Hence, we use it as a feature, which is represented as f_{SB}.

4 EXPERIMENTS

In this section, we discuss our experimental results. First, we present the IQ-Net model's performance with different encoders (CNN and BERT) and compare it with our baseline method. Then, we conduct an ablation study to understand the importance of each feature. Also, we conduct additional analysis over specific examples.

4.1 Datasets

We collect an annotated turn-level user perceived defect dataset for experiments by following the same annotation workflow as described in [4, 31]. We randomly sampled data for annotation. The dataset contains hundreds of thousands of samples.

The two examples for the first-turn with $label = 1$ and $label = 0$ are as follows.

Defect = 1 Example:
User request: Do I have any appointments today?
ICA response: Appointment is [definition of appointment]
User request: Tell me my appointments today
Defect = 0 Example:
User request: Turn on the lights
ICA response: Ok
User request: Thank you

4.2 Main Results

Table 4 illustrates the result of IQ-Net with two different encoders, including CNN and BERT, compared with result from baseline.

Perf(%)	Accuracy	F1	Recall	Precision
Baseline				
meta_data only	82.54	74.30	75.48	73.16
+ $< U_1, U_2 >$	+0.27	+0.64	+1.36	-0.03
+ $< U_1, R_1 >$	+0.22	+0.57	+1.3	-0.11
+ $< U_1, R_1, U_2 >$	+0.08	+0.73	+2.56	-0.92
IQ-Net (CNN)				
+ $< U_1, U_2 >$	+0.35	+1.21	+3.41	-0.75
+ $< U_1, R_1 >$	+1.82	+2.39	+1.43	+3.31
+ $< U_1, R_1, U_2 >$	+2.63	+3.74	+3.28	+4.18
IQ-Net (BERT)				
+ $< U_1, U_2 >$	+0.82	+1.09	+0.71	+1.44
+ $< U_1, R_1 >$	+2.69	+3.83	+3.38	+4.25
+ $< U_1, R_1, U_2 >$	**+3.23**	**+4.62**	**+4.14**	**+5.18**

Table 4: Results. Baseline: The baseline method is a Gradient Boosting Decision Tree (GBDT) [14] with the same features mentioned in Section 3. IQ-Net(CNN): our proposed method with CNN encoder. IQ-Net(BERT): our proposed method with BERT encoder.

As shown in Table 4, the IQ-Net (with either CNN encoder or BERT encoder) has better performance than the baseline method. IQ-Net (BERT) outperforms the baseline method of meta-data only with an improvement of 3.23% in accuracy, 4.62% in F1 score, and 5.18% in precision, and it outperforms the baseline method with full features with an improvement of 3.15% in accuracy, 3.89% in F1 score, and 6.1% in precision.

4.3 Ablation Study

Ablation for different features: We perform ablation experiments over the list of features used in IQ-Net(BERT) to better understand their relative importance. In Table 5, we show how much degraded the overall model performance is when we remove each specific feature. In particular, removing the context representation of $f_{<U_1,R_1,U_2>}$ will impact the overall performance the most, the decrease of accuracy is -3.237%, F1 score is -4.615%.

4.4 Case Analysis

As shown in Table 4, using both $< U_1, U_2 >$ and $< U_1, R_1 >$ as context helps the IQ-Net have better performance than considering only one of the signals. We look into examples where the former can make a correct prediction while the latter fails to do so. For the *defect = 1* example below, the predicted probability score of the second

Perf(%)	Accuracy	F1	Recall	Precision
IQ-Net(BERT)	**85.77**	**78.92**	**79.62**	**78.34**
$-f_{<U_1,U_2>}$	-0.542	-0.792	-0.758	-0.824
$-f_{<U_1,R_1>}$	-2.412	-3.530	-3.423	-3.632
$-f_{<U_1,R_1,U_2>}$	-3.237	-4.615	-4.131	-5.076
$-f_{BI}$	0.018	0.050	0.139	-0.036
$-f_{TM}$	-0.195	-0.076	0.727	-0.836
$-f_{GT}$	-0.888	-1.752	-3.255	-0.243
$-f_{UI}$	0.466	0.306	-1.172	1.787
$-f_{UD}$	-0.265	-0.215	0.445	-0.842
$-f_{DS}$	-0.350	-0.523	-0.547	-0.500
$-f_{PI}$	-0.074	-0.064	0.107	-0.228
$-f_{SB}$	-0.209	-0.221	0.109	-0.537

Table 5: Results of IQ-Net(BERT) on Feature Ablation.

turn being a rephrase of the first turn is 0.432 and the predicted relevance score between the request-response pair is 0.906. Thus, the defect example will not be easily captured if only considering one of the $< U_1, U_2 >$ or $< U_1, R_1 >$ pairs as context. However, the overall IQ-Net can detect it as a defect by considering both at the same time.

Defect = 1 Example:
User request: where is university located?
ICA response: University, Hillsborough County, ...
User request: where is yale university

5 CONCLUSION

In this paper, we propose to build an automated metric to evaluate dialogue quality at turn level for ICAs. We propose an IQ-Net model with end-to-end tuned from raw dialogue context and system metadata that allows us to predict interaction level dialogue quality. Experimental results show that our methods outperform the baseline method and work well across different domains as well as various intents. We conduct an ablation study on individual features to understand the contribution of each feature on model's prediction ability.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 19–26.

[2] Snjezana Babic, Tihomir Orehovacki, and Darko Etinger. 2018. Perceived user experience and performance of intelligent personal assistants employed in higher education settings. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 0830–0834.

[3] Ana Berdasco, Gustavo López, Ignacio Diaz, Luis Quesada, and Luis A Guerrero. 2019. User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana. In *Multidisciplinary Digital Publishing Institute Proceedings*, Vol. 31. 51.

[4] Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019. Multi-domain Conversation Quality Evaluation via User Satisfaction Estimation. *arXiv preprint arXiv:1911.08567* (2019).

[5] Thomas M Brill, Laura Munoz, and Richard J Miller. 2019. Siri, Alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications. *Journal of Marketing Management* 35, 15-16 (2019), 1401–1436.

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*. 737–744.

[7] Eunah Cho, He Xie, and William M Campbell. 2019. Paraphrase generation for semi-supervised learning in NLU. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. 45–54.

[8] Heriberto Cuayáhuitl, Seonghan Ryu, Donghyeon Lee, and Jihie Kim. 2018. A study on dialogue reward prediction for open-ended conversational agents. *arXiv preprint arXiv:1812.00350* (2018).

[9] Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. 2020. Intelligent Personal Assistants: A Systematic Literature Review. *Expert Systems with Applications* (2020), 113193.

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

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Layla El Asri, Hatim Khouzaimi, Romain Laroche, and Olivier Pietquin. 2014. Ordinal regression for interaction quality prediction. In *Proceedings of ICASSP* (proceedings of icassp ed.). https://www.microsoft.com/en-us/research/publication/ordinal-regression-interaction-quality-prediction-2/

[13] 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*. 170–177.

[14] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

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

[16] Ahmed Hassan and Ryen W White. 2013. Personalized models of search satisfaction. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2009–2018.

[17] David Herbert and Byeong Kang. 2019. Comparative Analysis of Intelligent Personal Agent Performance. In *Pacific Rim Knowledge Acquisition Workshop*. Springer, 127–141.

[18] 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 *International Workshop on Spoken Dialogue Systems Technology*. Springer, 48–60.

[19] Katja Hofmann, Lihong Li, Filip Radlinski, et al. 2016. Online evaluation for information retrieval. *Foundations and Trends® in Information Retrieval* 10, 1 (2016), 1–117.

[20] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[21] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web*. 506–516.

[22] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 45–54.

[23] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 121–130.

[24] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 113–122.

[25] Tihomir Orehovački, Snježana Babić, and Darko Etinger. 2018. Modelling the perceived pragmatic and hedonic quality of intelligent personal assistants. In *International Conference on Intelligent Human Systems Integration*. Springer, 589–594.

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[27] Pragaash Ponnusamy, Alireza Roshan Ghias, Chenlei Guo, and Ruhi Sarikaya. 2019. Feedback-Based Self-Learning in Large-Scale Conversational AI Agents. *arXiv preprint arXiv:1911.02557* (2019).

[28] Louisa Pragst, Stefan Ultes, and Wolfgang Minker. 2017. *Recurrent Neural Network Interaction Quality Estimation*. Springer Singapore, Singapore, 381–393. https://doi.org/10.1007/978-981-10-2585-3_31

[29] Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2017. Interaction Quality Estimation Using Long Short-Term Memories. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Saarbrücken, Germany, 164–169. https://doi.org/10.18653/v1/W17-5520

[30] Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. 2016. Prediction of prospective user engagement with intelligent assistants. In *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 1203–1212.

[31] Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. 2017. Predicting causes of reformulation in intelligent assistants. *arXiv preprint arXiv:1707.03968* (2017).

[32] Ruhi Sarikaya. 2017. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine* 34, 1 (2017), 67–81.

[33] 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 (2015), 12–36.

[34] 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 *LREC*. 3369–3373.

[35] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26, 3 (2000), 339–373.

[36] Pei-Hao Su, Milica Gašić, and Steve Young. 2018. Reward estimation for dialogue policy optimisation. *Computer Speech & Language* 51 (2018), 24–43.

[37] Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

[38] Stefan Ultes and Wolfgang Minker. 2014. Interaction Quality Estimation in Spoken Dialogue Systems Using Hybrid-HMMs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, Philadelphia, PA, U.S.A., 208–217. https://doi.org/10.3115/v1/W14-4328

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