DialogueScore: Evaluating Responses in Task-Oriented Dialogue

Anonymous ACL submission

Abstract

Task-Oriented Dialogue systems have been widely deployed in real-world applications in the last few years. Yet, evaluations of task-oriented dialogue systems are relatively limited. The informative and success score only consider the key entities in the generated responses to judge whether the user's goal is achieved. On the other hand, the fluency metric (BLEU score) cannot measure the quality of the short responses properly since the golden responses could be diversified. To better explore the behavior and evaluate the generation ability of task-oriented dialogue systems, we explore the relation between user utterances and system responses and their follow-up utterances. Therefore, we design a scorer named **DialogueScore** based on the natural language inference task and synthesize negative data to train the scorer. Via performances of **DialogueScore**, we observe that the dialogue system fails to generate high-quality responses compared with the reference responses. Therefore, our proposed scorer could provide a new perspective for future dialogue system evaluation and construction.

1 Introduction

Task-Oriented Dialogue system (Young et al., 2013) is a natural language processing task that aims at accomplishing user's goals. The process of task-oriented dialogue (TOD) usually consists dialogue state tracking, dialogue policy for making actions and response generation. Recent pretrained models directly generate system responses to user's utterances in a sequence-to-sequence generation manner. The process of evaluating the effectiveness of these dialogue systems is to discriminate whether the user's goals are accomplished by the generated responses (e.g. using the Informative score and Success score (Budzianowski et al., 2018)) and whether the generated responses are fluent (e.g. using BLEU score (Papineni et al., 2002) or ROUGE score (Lin, 2004)).

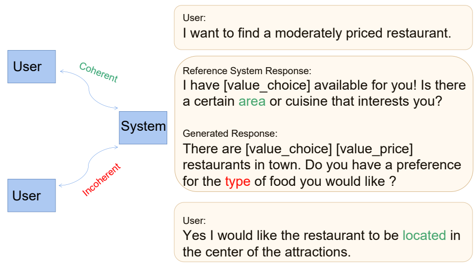

Figure 1: The utterance-response pair is coherent while the response and the follow-up utterance pair is not (ask for type of food but respond with location of food). The [value] tokens are later filled with knowledge base information given from the Dataset.

However, these existing metrics cannot fully measure the quality of the generated responses: the informative and success metrics only focus on whether the key entities (e.g. places to travel or restaurant to book) exist in the generated responses. That is, once the desired entities are detected in the generated responses, the session will be considered successfully responded. On the other hand, the fluency score such as BLEU is less effective since the dialogue between the user and system is usually short and precise about the goal in the dialogue session. In the task-oriented dialogue tasks, the evaluation process uses golden user utterances in previous turns and measure the quality of current turn responses regardless of the possibility that the generated responses could affect the user utterances in the next turn making the dialogue session *unnatural*. Human evaluation is often the best indicator of the effectiveness of deep learning systems. Human-involved evaluation in dialogue systems is difficult to construct since the dialogue scenario is hard to reconstruct, therefore automatic scorers are needed.

In this paper, we focus on scoring the quality of the generated responses in the TOD system and explore the relation between the responses and their corresponding dialogue sessions.

In task-oriented dialogue tasks, we assume that a good response should be coherent within the session even when the responses might not match the reference responses since the reference responses are static. We consider that by measuring the coherence between the responses and their corresponding dialogue sessions, we can measure the quality of the generated responses. The generated responses should match the user's utterances and it should also match the content of the follow-up utterances. Therefore, we introduce two types of scorer: (1) utterance and response matching scorer (2) response and follow-up utterance matching scorer as illustrated in Figure 1. With these two types of scorers, we can measure the coherence of the generated responses as an auxiliary tool when the BLEU score fails to measure the quality of the generated responses. Further, in task-oriented dialogue systems, the dialogue history is extremely important, we measure the current pair of user utterances of responses based on a certain number of previous turns (e.g. 1/2 turns of history).

To train the proposed scorer, we use annotated user-system conversation pairs as positive pairs and construct negative pairs based on synthetic data that is similar in dialogue history, topic and structure information. We construct negative data based on similar dialogue turns from several perspectives which contain session-similar, action/state similar and domain similar negatives. We train the scorer as a classification task using synthesized data and obtain the confidence of the classification results as the predicted score.

Experimentally, we train the scorer based on pre-trained models exemplified by BERT (Devlin et al., 2018) and test on state-of-the-art task-oriented dialogue systems. Through **DialogueScore** performance comparisons between the generated responses and reference responses, we observe that the state-of-the-art systems cannot generate responses that have similar DialogueScore performances compared with the reference responses, indicating that the scorer can serve as a valuable evaluator. We also conduct a human-involved meta-evaluation to score the coherence of the utterance-response pair to prove that the proposed score is similar to human judgments. Therefore, we believe that the proposed **DialogueScore** serves as an additional tool to evaluate responses in TOD and provides a new perspective to improve current dialogue systems.

2 Related Work

2.1 Task-Oriented Dialogue Systems

Task-Oriented Dialogue (TOD) task is a major dialogue task that aims at achieving user's goals such as booking flights, restaurants, etc.(Wen et al., 2016; Eric and Manning, 2017). We select the dialogue response generation scenario as our target task to evaluate.

Pipeline methods generate dialogue responses based on a natural language generation module while the dialogue state and policy are obtained by previous modules (Young et al., 2013).

With pre-trained models exemplified by BERT (Devlin et al., 2018), GPT (Radford et al., 2018), BART (Lewis et al., 2019) and T5 (Raffel et al., 2019), generating responses based on pre-trained sequence-to-sequence models (Yang et al., 2020; Hosseini-Asl et al., 2020; Su et al., 2021) is widely explored.

The evaluation metrics used in the task-oriented dialogue systems concern mainly two folds: (1) whether the user's goal is achieved; (2) quality of the generated responses. The basic metric is the informative score and the success score (Mehri et al., 2019) that measure whether the specific entities of user's goals (e.g. obtaining the address of the hotel, price of the restaurant) are all extracted and given in the responses. The response quality score is normally measured by BLEU score which uses in evaluating the fluency of generated texts in various tasks such as machine translation (Sutskever et al., 2014; Bahdanau et al., 2014). However, in dialogue tasks, BLEU score does not always indicate good quality responses since the generated texts are usually short and concise (Yang et al., 2020).

2.2 Neural Model Based Evaluation

Recent trends leverage neural models to automatically evaluate generated texts from different perspectives.

FactCC (Kryściński et al., 2019) introduces language inference systems to measure the factuality of the generated summarizations. The factuality checker is trained based on human-designed data that contains certain types of factual errors in the generated summarizations. The core idea of factuality checker is to construct a new task to learn the certain problem in the text generation process (e.g. factuality in the text summarization tasks), which is related to our work that explores the response closeness to the user utterances in task-oriented dialogue systems.

Dziri et al. (2019) introduces the NLI system to measure the response coherence (Dang, 2005) to the dialogue histories in open-domain dialogues tasks. Open-domain dialogue does not consider the dialogue states and actions which is different from task-oriented dialogue tasks. Therefore, Dziri et al. (2019) uses entailment-based models based on the MNLI data (Williams et al., 2018) without exploring the details between users and systems.

Recently, neural models can be used to construct automatic metrics. For instance, BERTScore (Zhang et al., 2019) uses token-level matching in the distributed representation space to measure the similarity of the generated texts and their references. BARTScore (Yuan et al., 2021) evaluates generated texts using the text generation model (e.g. BART (Lewis et al., 2019)). Methods such as BLEURT (Sellam et al., 2020) use a regression layer and train the scorer as a fine-tuning task based on the pre-trained models to imitate human judgments as well as metrics such as BLEU and ROUGE. The difference between constructing automatic metrics and building neural network based scorers is that metrics are calculated based on reference texts while scorers can give evaluation results independently.

3 DialogueScore

In this section, we first introduce the idea of scoring dialogue states and actions. Then we introduce the scorer that considers the scoring process as a natural language inference task. Finally, we introduce the construction and the training process of our scorer.

3.1 Scoring Response Quality

The basic idea of DialogueScore is to measure the quality of the generated system responses. In task-oriented dialogue systems, high quality responses should understand user's intentions and consist with the entire dialogue session. That is, the responses should be coherent within the entire dialogue session. Such quality cannot be properly measured by current evaluation strategies since the BLEU score focuses more on the faithfulness with the reference responses and Inform and Success score focus on key entities extraction. Specifically, in TOD, the responses are usually concise and faithful to achieving user's goals, the response quality

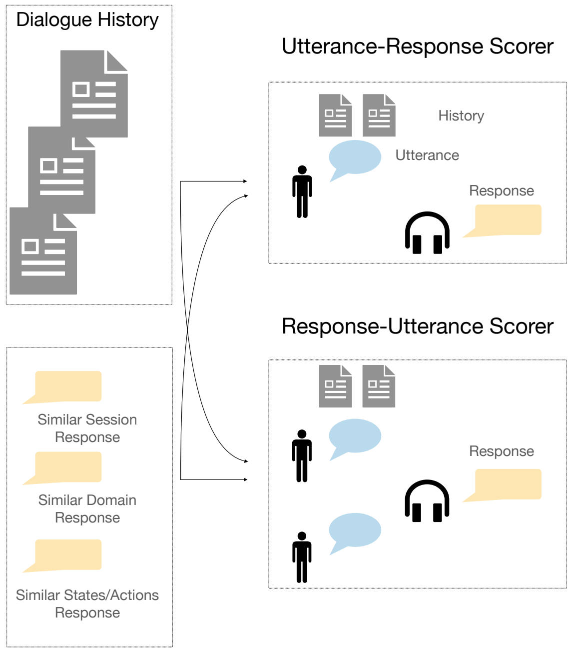

Figure 2: Process of DialogueScore. We first train two types of scorers with different number of dialogue histories, then we use these scorers to evaluate the system responses.

could be promising while the BLEU score is low since generated texts could be the paraphrasing of the reference texts.

Therefore, to measure the relation between generated responses and their surrounding texts within the dialogue session, we introduce two types of relation between the responses and the dialogue session. By exploring the connections between the responses and their corresponding user utterances and follow-up user utterances, we can measure the coherence of the responses within the dialogue session.

Utterance and Response Relation:

System responses are supposed to satisfy the user's queries, therefore, the system response should be coherent to the last user utterance. A proper response to the user utterance indicates that the system is familiar with the dialogue states in the current session.

Response and Utterance Relation:

Similarly, system responses should also be coherent to the follow-up user utterances. Human dialogue requires interactions between both customers and service providers, therefore, the generated responses should also focus on the coherence with the follow-up user utterances.

Considering these two types of relation, we build two scorers as an auxiliary evaluation tool to evaluate the quality of the generated responses when the responses are not matched with the reference responses.

3.2 Scoring with Dialogue Histories

The utterance-response scorer and the response-utterance scorer described above lacks the dialogue history information which is the major topic of dialogue tasks. Therefore, we improve our coherence scorer with dialogue histories.

Specifically, we construct scorers that concern different number of previous turns in predicting the coherence of the current pair. That is, we add dialogue histories in the front of the utterance-response pair for the scorer to learn information about previous dialogue sessions. Here, we do not consider all previous dialogue histories since we aim to fairly measure the response quality in different turns. Otherwise, responses from the last turns receive more dialogue history information which could bring extra information or noise to the scorer.

3.3 NLI System as Scorer

It is straightforward to incorporate natural language inference to construct a scorer to explore the relation between generated responses and their corresponding dialogue session since the NLI system is designed to measure the relation of a pair of texts.

Suppose we have T turns of dialogue in a given session, we define the t^{th} user utterance as $\boldsymbol{u}_t$. For the t^{th} system response to the user's queries, we define as $\boldsymbol{r}_t$. Besides current turns, we denote N turns of dialogue history before the t^{th} pair of utterance-response using $\boldsymbol{h_t} = [u_{t-N}, r_{t-N}, u_{t-N+1}, \cdots]$. Therefore, the NLI system $F(\cdot)$ that scores the consistency between the t^{th} response and utterance predicts $s = F([\boldsymbol{h_t}; \boldsymbol{u_t}], \boldsymbol{r_t})$ for utterance and response relation and $s = F([\boldsymbol{h_t}; \boldsymbol{u_t}, \boldsymbol{r_t}], \boldsymbol{u_{t+1}})$ for response and follow-up utterance relation. We use the *softmaxed* score as the final DialogueScore.

3.4 Data Synthesis

The scorer is supposed to understand whether the generated responses are natural in the dialogue contexts. We construct negatives that are similar to the reference responses as synthesized data to train the scorer.

We focus on different perspectives in the dialogue dataset:

- Session Information: We assume that information of responses in the same dialogue session is related. That is, the dialogue belief states and actions are consistent within the same session. Therefore, for the t^{th} pair, we use a randomly selected response $\boldsymbol{r}_s$ where $s \neq t$ from the same session as the negative response to construct the negative sample.

- Domain Information: In task-oriented dialogue systems, dialogue sessions from the same domain (e.g. booking flights or hotels) are supposed to share similar dialogue structures. That is, the procedure of booking a flight or finding a place to dine is similar among different sessions. Therefore, we introduce a simple negative construction method. We use the same order of system response $\boldsymbol{r^d}_t$ where $\boldsymbol{r^d}$ is the response from another same-domain session as the negative response of the current pair. That is, we use the response from the t^{th} turn in session $\boldsymbol{r}^d$ as the negative sample for the t^{th} turn in session $\boldsymbol{r}$.

- Dialogue States/Action Information: Besides the domain and session level similar information, we could directly use the dialogue actions and states information to construct negative responses. We use responses that share the same dialogue states and actions as negative responses.

We introduce several straightforward strategies to synthesize data to train the scorer that is able to evaluate the quality of the generated responses. Different from synthesizing data in building factuality checker in summarization evaluations, in task-oriented dialogue systems, the unnatural responses are relatively difficult to define. Therefore, instead of synthesizing certain types of unrelated responses or useless responses, we synthesize negatives by categories. We aim at certain perspectives which could be essential to the task-oriented dialogue systems. In task-oriented dialogue systems, information such as domain difference or action difference could significantly affect the response generation.

With our proposed scorer training process, we are able to construct multiple scorers that can be used in evaluating the quality of generated texts instead of simply counting key entities and using co-occurrence compared with the reference responses as evaluation.

4 Experiments

In this section, we construct experiments on task-oriented dialogue tasks using our proposed DialogueScore to evaluate the strong baselines and explore whether the state-of-the-art models can achieve satisfactory results on DialogueScore. Plus, we design experiments exploring the faithfulness of our proposed score compared with human judges.

4.1 Datasets

We use the Multi-WOZ 2.0 dataset (Budzianowski et al., 2018) which is the most widely used dataset in the task-oriented dialogue tasks. Multi-WOZ dataset contains end-to-end dialogue modeling, dialogue state tracking and user intent classification sub tasks. We only test the dialogue response modeling task since DialogueScore aims to evaluate the quality of the generated responses. In the Multi-WOZ dataset, the response generation process is constructed on the database state which is automatically retrieved from a database pre-constructed. We also use this pre-defined database information in the response generation. Further, we establish both full-data and few-shot settings based on the Multi-WOZ dataset. That is, we use a small proportion of the training set (10 %) to train the model.

4.2 Baseline Models

We use several state-of-the-art models and use our proposed DialogueScore to evaluate the quality of their generated responses.

The first model we test is the state-of-the-art model named PPTOD (Su et al., 2021) which includes a unique further pre-training stage initialized from the T5 (Raffel et al., 2019) model. The PPTOD model achieves the state-of-the-art performances on task-oriented dialogue tasks using text-to-text pre-training task that specially designed for task-oriented dialogue tasks.

We use the small version of the PPTOD model which contains 6 layers of encoder and 6 layers of decoder which has similar parameter number compared with BERT-base model.

We also test the UBAR (Yang et al., 2020) model which uses GPT model as their backbone model. The UBAR model uses dialogue states in the response generation process different from the vanilla sequence-to-sequence generation process used in the PPTOD model.

4.3 Scorer Training Details

We train our scorer based on the BERT-base-uncased model (Devlin et al., 2018) following the implementations provided by Huggingface Transformers (Wolf et al., 2019). Following the details of fine-tuning NLI tasks, we train the scorer with hyper-parameters listed below. We set batch size to 64 with learning rate 2e-5 and run 3 epochs on NVIDIA 3090 GPUs.

The numbers of using more number of history turns in the synthesized dataset are smaller since some sessions do not have enough turns to construct dialogue histories.

4.4 Metrics

4.4.1 Automatic Evaluation

The traditional evaluation metric includes Inform, Success and BLEU score. Following , a combined score is introduced: Combined = (Inform + Success) / 2 +BLEU.

In the Inform and Success calculation, the score is calculated based on the entire session.

In the BLEU score calculation, we calculate the BLEU score based on the reference responses and the generated responses. The evaluation unit is one turn in a dialogue session.

Our proposed DialogueScore measures the turn-level response quality therefore the evaluation unit is also a turn. We calculate the average score of multiple turn history DialogueScore as the final score measuring the utterance and response coherence and the response and follow-up utterance coherence.

4.4.2 Human Evaluation

Besides DialogueScore evaluation and traditional metrics evaluation, we introduce a human-involved meta evaluation to measure the quality of the generated responses. Through this meta evaluation, we are able to calculate the correlation coefficient between the evaluation scores and the human ratings. Specifically, we select 50 dialogue responses and give certain number (4 turns) of dialogue histories plus the follow-up utterance and ask human judges to score whether the responses are natural in the given dialogue session.

Following Su et al. (2021), we ask multiple human judges to evaluate the responses withe respect to whether the responses are coherent within the session and use the averaged score. Human judges need to predict the responses as

Model / Metric	Reference	PPTOD	UBAR	PPTOD-fewshot
Traditional Metrics				
Inform	-	87.8	85.1	83.5
Success	-	75.3	71.0	68.2
BLEU	-	19.9	16.2	15.6
DialogueScore: Utterance-Response				
0-His.	53.5	56.8 ↑3.3	56.4↑2.9	58.0↑4.5
1-His.	54.2	56.7↑2.5	54.2↑0	58.5↑4.3
2-His.	52.8	53.9↑1.1	52.1↓0.7	59.1↑6.3
3-His.	51.3	43.1↓8.2	29.0↓22.3	36.3↓15
4-His.	51.5	44.4↓7.1	33.0↓18.5	43.4↓8.1
DialogueScore: Response-Utterance				
0-His.	55.0	34.4↓20.6	35.6↓19.4	33.5↓21.5
1-His.	53.7	33.1↓20.6	32.2↓21.5	30.9↓22.8
2-His.	53.2	34.9↓18.3	30.9↓22.3	31.0↓22.2
3-His.	49.6	42.3↓7.3	32.2↓17.4	32.5↓17.1
4-His.	50.0	44.3↓5.7	39.5↓10.5	41.6↓8.4

Table 1: Model Evaluation with DialogueScore on the Testset of Multi-WOZ 2.0 Dataset. The arrow suggests the gap between the generated and reference texts.

fully(2)/partial(1)/zero(0) related to the session based on the dialogue histories and the follow-up user utterance. We then calculate the correlation coefficient between scores (BLEU and DialogueScore) and the meta evaluation results.

4.5 Results

In Table 1, we show the evaluation results of the DialogueScore on state-of-the-art models as long as the traditional metric results. In Table 2, we show the evaluation results of DialogueScore metric measured by the Kendall and Spearman correlation index compared with the meta evaluation.

4.5.1 Response Evaluation with DialogueScore

As seen in Table 1, state-of-the-art models are struggling in achieving promising results of DialogueScore.

In Utterance-Response DialogueScore which considers over 3 turns of dialogue histories, the generated texts from the best model PPTOD are still 7 points worse than the reference responses, indicating that the generated texts are less satisfactory when the scorer considers multiple turns of dialogue histories. The generated responses do not match their corresponding user utterances, indicating that these responses do not understand the dialogue states given previous dialogue information.

Also, we can observe that when the dialogue history number is small (less than 3 turns), the generated texts can obtain higher performances than the reference responses in the utterance-response score. We assume that the neural dialogue systems always respond to the user's queries without deeply understanding the dialogue states. Therefore, when the scorer only considers the utterance and response pair without any dialogue histories, the query-answer pattern is prevailing and the generated responses are more concise than reference responses. On the other hand, the reference responses are sometimes spontaneous which makes it harder for automatic scorers to evaluate. We explore this phenomenon in detail in the later section.

Further, in the results of the Response-Utterance DialogueScore, the performances of the generated texts are constantly worse than the reference responses. The generated responses cannot match the follow-up user utterances, indicating that these generated responses cannot anticipate the user's goals properly. The generated responses only answer to the user's queries without paying attention to the entire dialogue session, which is more similar to a question-answering system instead of dialogue system.

4.5.2 Model-Wise Evaluation

In Table 1, we observed that different models perform differently in both traditional metrics and DialogueScore. Compared with standard PPTOD model, the fewshot PPTOD model that uses only 10 % training data achieves unsatisfied performances in DialogueScore. We can notice that when the data is limited, the model focus more on the utterance-response pair without considering dialogue histories, which results in the highest performances in the 0-turn and 1-turn history utterance-response scorer. On the other hand, we can observe that when the traditional performances are similar, the DialogueScore could show a larger difference between the PPTOD model and its previous baseline UBAR model.

4.5.3 Meta Evaluation of DialogueScore

We calculate the correlation coefficient score between each automatic evaluation metrics and the the meta evaluation score. Further, we consider that DialogueScore can fairly score generated texts that are not match the reference texts by the BLEU score, therefore we construct an additional testset that only selects low BLEU score (lower than 1.0 BLEU score) samples from the testset annotated by human judges.

As seen in Table 2, we can observe that:

Model Metric	Average	Kendall	Spearman
Traditional Metrics			
Inform	87.7	-	-
Success	75.0	-	-
BLEU	18.9	0.17	0.24
BLEU(lowBLEU)	0.5	0.20	0.27
DialogueScore: Utterance-Response			
0-His.	57.0	0.13 $\downarrow$0.04	0.14 $\downarrow$0.10
1-His.	56.8	0.19$\uparrow$0.02	0.21$\downarrow$0.03
2-His.	53.9	0.25$\uparrow$0.08	0.28$\uparrow$0.04
3-His.	43.0	0.09$\downarrow$0.08	0.10$\downarrow$0.14
4-His.	44.4	0.06$\downarrow$0.11	0.08$\downarrow$0.16
DialogueScore: Response-Utterance			
0-His.	34.5	0.53$\uparrow$0.36	0.61$\uparrow$0.37
1-His.	33.4	0.48$\uparrow$0.31	0.56$\uparrow$0.32
2-His.	35.3	0.28$\uparrow$0.11	0.33$\uparrow$0.07
3-His.	42.4	0.24$\uparrow$0.07	0.30$\uparrow$0.06
4-His.	44.3	0.37$\uparrow$0.20	0.38$\uparrow$0.14
Utterance-Response(lowBLEU)			
0-His.	52.9	0.34$\uparrow$0.14	0.38$\uparrow$0.11
1-His.	56.8	0.24$\uparrow$0.04	0.27$\uparrow$0.01
2-His.	45.7	0.27$\uparrow$0.07	0.31$\uparrow$0.04
3-His.	39.9	0.08$\downarrow$0.12	0.08$\downarrow$0.19
4-His.	44.0	0.05$\downarrow$0.15	0.10$\downarrow$0.17
Response-Utterance(lowBLEU)			
0-His.	32.2	0.38$\uparrow$0.18	0.45$\uparrow$0.18
1-His.	35.6	0.40$\uparrow$0.20	0.48$\uparrow$0.21
2-His.	41.6	0.30$\uparrow$0.10	0.34$\uparrow$0.07
3-His.	46.7	0.25$\uparrow$0.05	0.28$\uparrow$0.01
4-His.	45.7	0.28$\uparrow$0.08	0.32$\uparrow$0.05
Human Rating	1.10	-	-
Human Rating (lowBLEU)	1.00	-	-

Table 2: Metric evaluation of DialogueScore. The arrow suggests the gap between DialogueScore and the corresponding BLEU score.

Model Metric	DS	Session	Domain	State/Act.
DialogueScore: Utterance-Response				
0-His.	53.5/56.8	82.4/86.6	72.2/76.0	62.2/65.5
1-His.	54.2/56.7	85.9/88.7	75.1/79.2	63.4/65.2
2-His.	52.8/53.9	83.4/81.3	76.6/80.0	62.6/66.3
3-His.	51.3/44.4	84.3/82.2	78.2/78.9	63.1/58.9
4-His.	55.0/34.4	83.4/75.9	80.0/73.0	62.9/56.8
DialogueScore: Response-Utterance				
0-His.	55.0/34.4	84.5/65.0	73.0/52.5	64.8/48.6
1-His.	53.7/33.1	84.9/67.0	73.6/53.1	65.1/51.1
2-His.	53.2/34.9	83.0/64.7	73.7/56.2	64.2/53.3
3-His.	49.5/42.3	81.0/74.3	74.8/56.3	65.1/57.0
4-His.	50.0/48.9	80.4/73.6	74.4/55.3	67.5/60.0

Table 3: Ablation Studies of scorers using different synthetic data on the Multi-WOZ 2.0 testset using the PPTOD-small model. -/- is the reference response score and PPTOD model output score correspondingly.

High Correlation in DialogueScore: in the response and follow-up utterance scorer, we can observe that the scorer considers limited dialogue histories obtain high correlation score with human judges, indicating that the scorer can successfully tell whether the generated responses are coherent within the dialogue session.

Scorer with Dialogue Histories: we found that with limited history or too much dialogue histories, the user-response scorer does not show significant correlation with human judges while show significant correlation with reference responses. This unusual phenomenon indicates that in TOD systems, the response answered by human might not consist with other human judges considering that the dialogue sessions cannot be easily re-constructed. Therefore, strong automatic scorers are needed in TOD system to imitate human behaviors since human responses and judgements might be different.

Low-BLEU Score Samples:

We introduced DialogueScore as an auxiliary tool to score the quality of the generated response. Therefore, we establish an experiment to explore cases where the BLEU score fails to measure the response quality. We select samples with low bleu score and calculate the corresponding correlation score between DialogueScore and human judges. As seen, the human ratings drop by 0.1 point while DialogueScore shows significant improvement compared with the BLEU score, indicating that when the BLEU score becomes less effective, the DialogueScore can serve as an additional tool to help evaluate the response quality.

4.6 Ablations

We conduct experiments to explore the construction process of our proposed scorer. That is, we only consider a certain type of synthesized data in training the scorer. We test the reference response DialogueScore and the generated response DialogueScore and show in Table 3. As seen, when the scorer is trained on one type of synthesized negative data, the average score is higher than the scorer trained by all synthesized data combined. We can observe that scorer trained using the state and action similar negatives does not show significant difference between reference and generated responses. This indicates that though the purpose of the scorer is to evaluate the quality of the generated responses, directly using negatives with similar states/actions does not help much. Using different perspectives of negatives can help the scorer to understand better about the quality of the generated responses. Therefore, we believe that a stronger scorer trained on higher quality negatives could obtain better results in response evaluation.

Type	Text	BLEU	DialogueScore
Reference Dialogue Session	User: hi, i need a train to broxbourne , please. System: what day will you be traveling?	-	0.98/0.99
	User: i will be leaving on saturday and i am departing from cambridge . System: there are [value_choice] trains that day do you have a time you will be traveling? User: i need to arrive by 16:15 and i want to book it for 6 people .	-	0.96/0.97
Generated Dialogue Session	User: hi , i need a train to broxbourne , please . System: i can help you with that . where will you be departing from ?	0.43	0.98/0.99
	User: i will be leaving on saturday and i am departing from cambridge . System: there are [value_choice] trains leaving [value_departure] on [value_day] . what time would you like to leave ? User: i need to arrive by 16:15 and i want to book it for 6 people .	16.46	0.96/0.97

Table 4: Difference in the BLEU score and Dialogue Score evaluation.

4.7 Analysis

4.7.1 Difference between BLEU and DialogueScore

The major problem about using the BLEU score in evaluating TOD systems is that the BLEU score highly rely on the co-occurrence between the reference response and the generated response. While in response generation, texts can be very different with the same purpose in TOD systems. The case study in Table 4 also proves such a phenomenon: the responses generated by the neural networks can understand the user's goals and give proper suggestions for the user. Yet the representation does not match the reference responses which results in poor BLEU score. In such a case, we can observe that the DialogueScore can give high confidence that the generated texts are fluent and proper as system responses.

Therefore, the matching of tokens might not be necessary when it comes to system responses in the TOD system. We can conclude that in the TOD systems, automatic scorer can be helpful in improving dialogue systems as a proper evaluation guidance. Our proposed DialogueScore could be a possible direction of exploring automatic evaluations in the dialogue systems.

4.7.2 Variance in Multi-Turn DialogueScore

As illustrated above, in Table 1, the generated responses achieve even higher performances than the reference responses in the utterance-response DialogueScore with small number of histories (smaller than 3 turns).

When the utterance-response scorer only considers limited number of dialogue histories, the scorer can only focus on the query-answer pair between the user and the system.

As seen in the case study in Figure 3, when the scorer is able to consider previous histories, it can understand that the question asked by the user

Figure 3: Case Study of Multi-turn DialogueScore. The 0-turn-history DialogueScore focuses on the query asked by the user therefore predict a relatively lower score. When the scorer focuses on more dialogue history, the scorer can give higher confidence about the response quality.

in the follow-up utterance is a proper response to the system question "can I help you with anything else?". Without considering the dialogue histories, the scorer does not understand the user's intent about looking for the attraction, therefore a follow-up question may be considered inconsistent with the system response. Therefore, it is effective that we construct multiple scorers which consider different numbers of dialogue histories.

5 Conclusion

In this paper, we explore the possibility of better evaluation of generated responses in TOD systems. We propose an automatic scorer **DialogueScore** to measure responses based on not only previous user utterances but follow-up user utterances. We construct experiments to show that the generated responses are less satisfactory evaluated by DialogueScore. We are hoping that such a scorer can provide a potential direction in building task-oriented dialogue systems.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. *arXiv preprint arXiv:1904.03371*.

Mihail Eric and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog. *arXiv preprint arXiv:1907.10016*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *CoRR*, abs/2109.14739.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2020. Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2. *arXiv preprint arXiv:2012.03539*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.