Rethinking Response Evaluation from Interlocutor's Eye for Open-Domain Dialogue Systems

Yuma Tsuta[1]   Naoki Yoshinaga[2]   Shoetsu Sato[2*]   Masashi Toyoda[2]

[1]The University of Tokyo

[2]Institute of Industrial Science , The University of Tokyo

`{tsuta,ynaga,shoetsu,toyoda}@tkl.iis.u-tokyo.ac.jp`

Abstract

Open-domain dialogue systems have started to engage in continuous conversations with humans. Those dialogue systems are required to be adjusted to the human interlocutor and evaluated in terms of their perspective. However, it is questionable whether the current automatic evaluation methods can approximate the interlocutor's judgments. In this study, we analyzed and examined what features are needed in an automatic response evaluator from the interlocutor's perspective. The first experiment on the Hazumi dataset revealed that interlocutor awareness plays a critical role in making automatic response evaluation correlate with the interlocutor's judgments. The second experiment using massive conversations on X (formerly Twitter) confirmed that dialogue continuity prediction can train an interlocutor-aware response evaluator without human feedback while revealing the difficulty in evaluating generated responses compared to human responses.

1   Introduction

Along with the growth of open-domain dialogue systems (Xu et al., 2022b,c; Bae et al., 2022; Takasaki et al., 2023), it is crucial to develop automatic methods that efficiently evaluate those systems. The automatic evaluations usually qualify system responses for utterances sampled from human conversation logs (§ 2). Since Liu et al. (2016) showed that automatic evaluation with a single reference response such as BLEU (Papineni et al., 2002; Forgues et al., 2014) did not correlate with human judgments due to the response diversity in open-domain dialogue (Sato et al., 2017; Tsuta et al., 2020), unsupervised reference-free methods and supervised methods that mimic human judgments have become popular (Yeh et al., 2021). However, these studies evaluate their methods in

Figure 1: A discrepancy between interlocutor and outsider evaluations for open-domain dialogue systems.

terms of correlation with judges by third-party annotators (*outsiders*), not partaking in the dialogue.

Do the existing methods correctly evaluate the dialogue systems? As illustrated in Figure 1, the interlocutor and evaluators may prefer different yet valid responses. Although Ghazarian et al. (2022) experimentally confirmed a poor correlation with outsider and interlocutor evaluations in terms of appropriateness, they remain focused on outsider evaluations. This study focuses on interlocutor evaluations to enable an automatic evaluation from the interlocutor's perspective. In the experiment, we concentrate on validating our ideas in terms of engagement. Because this metric is more subjective and varies across people.

In this study, for estimating the interlocutor's evaluations, we first analyze the effectiveness of personalizing the evaluation model to the target interlocutor. This is inspired by research on response generation (Li et al., 2016; Xu et al., 2022b), as it has been reported to be important to adjust (personalize) utterances to the interlocutor. For this analysis, we used the Hazumi dataset (Komatani and Okada, 2021), and confirmed that, even when we train a supervised evaluator to mimic interlocutor scores, it cannot accurately predict their scores without making it aware of the target interlocutor.

Motivated by the lessons learned from the above experiments, we then explore automatic response evaluation from the interlocutor's eye (§ 4). To reduce the cost of annotation, we utilize a dialogue continuity prediction (DCP) task to train an
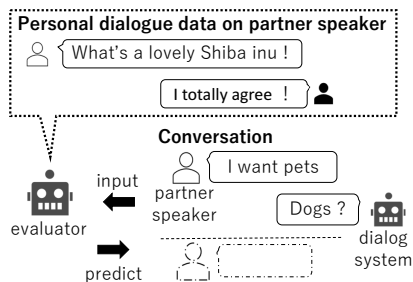
Figure 2: Automatic response evaluation via dialogue continuity prediction from the interlocutor's perspective.

interlocutor-aware evaluator (Figure 2). This task of estimating whether the target speaker will continue speaking or not can take advantage of labels (conversation stop signals) that are naturally annotated by the interlocutor in the conversation log. Experimental results on a conversation log on X (formerly Twitter) confirmed that the interlocutor-aware evaluator can be learned through the DCP task without human feedback while revealing the challenge of evaluating the system responses.

2 Related work

Automatic evaluation of dialogue systems To efficiently develop open-domain dialogue systems, researchers have sought evaluation methods that correlate with human evaluations. Since Liu et al. (2016) showed that reference-based metrics (Papineni et al., 2002; Forgues et al., 2014) using single reference responses do not correlate with human judgments, some studies use multiple reference responses (Galley et al., 2015; Gupta et al., 2019; Tsuta et al., 2020), while others train models by referring to human judgments (Lowe et al., 2017; Ghazarian et al., 2020) or other cues indicating valid responses (Tao et al., 2018; Ghazarian et al., 2019; Gao et al., 2020; Mehri and Eskenazi, 2020b; Xu et al., 2022a; Ghazarian et al., 2022). Recent studies (Mehri and Eskenazi, 2020a; Zhang et al., 2021) rely on language comprehension skills of pre-trained language models such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019). There are evaluation tasks from the other perspective such as dialogue breakdown detection (Higashinaka et al., 2016). The above studies were, however, developed to follow outsider evaluations and do not assume evaluation from the interlocutor's eye, even though some recent dialogue systems are being adapted to interlocutors in long-term conversations (Xu et al., 2022b,c; Bae et al., 2022; Takasaki et al., 2023).

A few studies have elucidated the relationship between user personality and the performance of dialogue systems from a psychological perspective (Guo et al., 2021; Papangelis et al., 2022). These studies suggest the importance of the interlocutor's traits in evaluating dialogue systems.

User-oriented NLP tasks There are several user-oriented (or personalized) NLP tasks in which users prefer different outputs and hence the systems are expected to be adjusted to match user preferences, including hashtag recommendations on social networking sites (Kywe et al., 2012) and website recommendations (Mishra et al., 2015). Similarly, for text generation tasks in which models have become able to generate decent outputs, researchers are starting to adapt the models to reflect individual preferences; examples of such tasks include summarization (Díaz and Gervás, 2007), machine translation (Mirkin and Meunier, 2015), text simplification (Bingel et al., 2018), and dialogue systems (Liu et al., 2020; Cho et al., 2022). To evaluate these systems, they need human judgments by the system users, which low reproducibility prevents us from efficiently developing the systems.

3 What is important to predict interlocutor evaluations?

To analyze what features are important for predicting interlocutor scores, we train score prediction models with several settings and compare their performances. Specifically, we analyzed the effect of reference scores (*e.g.*, interlocutor or outsider scores) and interlocutor-aware personalization on the evaluation models. Although Ghazarian et al. (2022) confirmed a low correlation between interlocutor and outsider evaluations, we further confirmed that outsider evaluations do not help predict interlocutor scores. For this analysis, we used the Hazumi dataset (Komatani and Okada, 2021), which is an open-domain conversation in the form of the Wizard of Oz experiment.

3.1 Hazumi dialogue datasets

For this analysis, we need a dataset that contains interlocutor and outsider scores to train and test models, and we utilize Hazumi1902 and Hazumi1911 subsets from the Hazumi dataset[1]. This dataset is an open-domain conversation in which "Wizard" behaves like a dialogue system and "Participant"

---

[1] https://www.nii.ac.jp/dsc/idr/en/rdata/Hazumi/

56

Dataset	Hazumi1902	Hazumi1911
# dialogues (participants)	30	30
Total count of exchanges	2477	2824
Utterance length (Wizard)	22.7	20.8
Utterance length (Participant)	22.2	25.1

Table 1: Statistics of the subsets of the Hazumi datasets after preprocessing (§ 3.1). Utterance length refers to the average number of characters in an utterance.

Training score	Target awareness	Pearson's r
Outsider		0.141
Outsider	✓	0.142
Interlocutor		0.166
Interlocutor	✓	**0.496**

Table 2: Results on interlocutor score prediction with ablation of training score and target speaker awareness.

speaks as the user. These subsets only contain an interlocutor's (*e.g.*, Participant's) and five outsiders' scores for each utterance by the Wizard. The participants and five outsiders rated the Wizard's utterances on a scale of 1 (feeling negative) to 7 (feeling positive) on the basis of user impressions.[2] The direction of the guideline is similar to the engagement metric in Ghazarian et al. (2020) and the annotation on the experiment in § 4.2 in terms of the willingness of dialogue continuity.

In what follows, we preprocess the dataset so that the exchanges, a pair of utterances by a Wizard and the Participant, consist of no empty utterance. After these preprocessing steps, we obtained 5301 exchanges from 60 dialogues. The detailed statistics are shown in Table 1. We split each conversation into 8:1:1 size chunks according to the flow of the conversation (and recombined) to train, validate, and test the prediction models.

3.2 Analyze the effective cues in interlocutor score prediction

We train evaluators using various cues to identify the interlocutor scores and clarify the requisite for automatic interlocutor evaluation. In this task, the models predict the interlocutor score to an utterance by Wizard. We feed Wizard's utterance and the longest contexts possible to the model, adding a special speaker token (`[Wizard]` or `[Participant]`) to distinguish who speaks utterances before the corresponding utterances.

Models We compared four evaluator models based on BERT (Devlin et al., 2019) for ablation. The differences between these models are i) whether to use interlocutor scores or (the averaged) outsider scores as the reference in training and ii) whether to use a speaker token specific to the target participant or the generic participant to-

ken preceding each utterance. The distinguished participant token is meant to adjust the evaluator to individual interlocutors, inspired by the speaker token introduced by Li et al. (2016) to model speakers in response generation.

Settings We fine-tuned each model from pre-trained Japanese BERT[3] for 10 epochs with the mean squared error loss. Other settings for the model were as follows: learning rate was $3e-5$ and optimized with AdamW (Kingma and Ba, 2015), and batch size was 64. We stored the model after each epoch and adopted the model that achieved the lowest loss for the validation data for testing.

Results Table 2 shows the correlations between model predictions and interlocutor actual scores. When a model is trained to predict the averaged outsider score, the evaluator showed a very low correlation of about 0.14. This confirms that outsider scores are useless in predicting interlocutor scores. Meanwhile, the model exhibits a much higher correlation when trained to predict interlocutor scores only with the awareness of target interlocutors; otherwise, the model shows only a slight improvement over the model learned by the averaged outsider scores. These results suggest that automatic interlocutor evaluation requires us to not only take the interlocutors' view (here, scores) into account but also to be aware of the target interlocutor.

4 Towards Automatic Response Evaluation from Interlocutor's Eye

From the result in § 3, we confirmed that accurate interlocutor score prediction requires personalizing the evaluator to the target interlocutor as well as referring to interlocutor scores. In practice, however, collecting interlocutor scores and creating conversations for the annotation are costly.

Therefore, focusing on evaluating responses in terms of engagement, we propose an alternative

---

[2]The annotation guidelines for user impressions define keywords and keyphrases such as "wants to keep talking" and "satisfied" as positive impressions, and "doesn't want to keep talking", "frustrated" and "confused" as negative impressions.

[3]https://huggingface.co/cl-tohoku/bert-base-japanese-v2

method to train an interlocutor-aware response evaluator via a dialogue continuity prediction task, assuming that utterances replied to by the interlocutors are more engaging than utterances without a response. The task is to predict whether there will be a response to an utterance in dialogue.[4]

4.1 Interlocutor Evaluation via Personalized Dialogue Continuity Prediction (DCP)

We train an automatic response evaluator via the dialogue continuity prediction task (Figure 2). The task settings are as follows. The task input is a conversation containing N utterances $U = \{u_0, u_1, ..., u_{N-1}\}$ made by two speakers s_i and s_j (u_{N-1} is made by s_j). The model output is assumed as the probability of whether the next response u_N is made by s_i, $P(u_N = exists \mid U, s_i)$.

How to consider the interlocutor in a model? As we have observed in § 3.2, it is crucial to personalize a response evaluator to the target interlocutor to estimate human judgments given by the interlocutors. Inspired by existing studies on personalizing open-domain dialogue systems (Li et al., 2016; Zhang et al., 2018), we consider two methods for the evaluator to take the interlocutor into account. The first method leverages a speaker token specific to the target interlocutor, which has been used in the experiments in § 3.2, whereas the second method refers to a user profile of the interlocutor. When we train a speaker token specific to the target interlocutor, we follow the procedure described in § 3.2. When using the profile, we input the profile text that accompanies the evaluation datasets (§ 4.2) at the beginning of the model inputs. We also consider the combination of a speaker-specific token and profile. In summary, we use three methods to model the interlocutor: using a speaker-specific token, using the profile, and using both methods simultaneously.

4.2 Experimental Setup

To investigate the effectiveness of our interlocutor-aware evaluators, we conduct experiments focusing on two metrics: 1) **accuracy of the dialogue continuation task** and 2) **correlation with manually-annotated engagement scores**.

Data Type	Train	Dev.	Test
Avg. turns in dialogue	3.4	3.4	3.3
Avg. char. size in turn	31.0	31.1	30.6
Avg. char. size in dialogue	106.5	106.8	101.2
Replied response size	1,779,895	100,899	1,088,970
No replied response size	1,244,530	70,135	832,377

Table 3: Statistics of the X dialogue datasets.

X (formerly Twitter) dialogue dataset We conducted the experiments using conversation logs on X. We can identify the author of a post, and handle a variety of users. We developed a Japanese dialogue dataset between two users using the API[5]. During the construction, we excluded posts that could be noisy, such as repetitive posts by bots, and preprocessed posts referencing studies using dialogues on Twitter (Li et al., 2016; Tsuta et al., 2020). In addition, we used only the conversations where all responses were made within 30 minutes because response rates tend to decrease over time (Gao et al., 2020). We expect these processes to make conversations more engaging, coherent, and less interrupted by others.

We randomly select 10,000 users who have had at least 30 conversations between January 2017 and March 2018. We use up to 400 conversations per user and their profile text to train the evaluator models. The profiles are collected with a field of the API (user.fields=description) and the average character size is 75.0. We used conversations of these users between March and December 2018 as test data. Because the intermediate reply is a positive sample and the last reply is a negative sample in the DCP task, several samples are collected from one conversation.

For the second experiment, we need conversations between a human (interlocutor) and a dialogue system, and the interlocutor's engagement score of willingness to reply to the system responses. Thus, we collected personal conversations on X by two members of our research group (a co-author and a graduate student) using the above same process. The dialogue data was added to the above dataset for (19, 6, and 10) and (165, 43, and 27) conversations as training, validation, and test data, respectively. Table 3 shows the statistics of the entire dataset.[6]

Dialog systems   To obtain system responses for human annotation, we employed seven dialogue models with two types of base architectures, Transformer encoder-decoder and decoder-only Transformer (GPT-2). As the encoder-decoder model, we used three publicly available dialogue systems that were trained with different datasets (Sugiyama et al., 2021).[7] As GPT-2, We fine-tuned a pretrained GPT-2[8] (medium) with our dataset (§ 4.2). We prepared four variations of fine-tuned GPT-2 to obtain dialogue systems with diverse conversation abilities. The two options are i) whether to re-initialize the model's parameters before fine-tuning and ii) whether to personalize the system to the interlocutor using a speaker token (Li et al., 2016).

Annotation with interlocutor judgments   To obtain manually annotated scores to responses for the second experiment, we asked the two annotators (same as the two interlocutors) to score seven responses generated by the above dialogue systems and one ground-truth response in the test data on a scale of 0 to 100, referring to Ji et al. (2022). 0 means that the annotator never responds to the last utterance of the conversation, and 100 means the opposite. We compensated the annotators at the rate of 1,050 JPY per hour.

Evaluator and baselines   We compare the following evaluation models. Because we also evaluate actual human responses, we use reference-free evaluation models that are easily available in our Japanese corpus as baseline models: BERT-NSP (Devlin et al., 2019)[9], BERT-RUBER (Ghazarian et al., 2019), FED (Mehri and Eskenazi, 2020a)[10] and Deep-AM-FM (Zhang et al., 2021). We also adopt the simple baseline model that always outputs the majority class label (i.e., whether or not to reply) based on the training data. We prepared two types of majorities: all users' majority (**Global majority**) and each interlocutor's majority (**Private majority**).[11]

For the baseline models, we adopted a pretrained BERT[3] for BERT-* and Deep-AM, and

---

[9] We employed the next sequence prediction task as an automatic evaluation model as in other studies (Mehri and Eskenazi, 2020b; Phy et al., 2020).

[10] We translated the follow-up utterances to evaluate system responses in terms of engagement.

[11] Note that we use this model only in the first experiment because the second experiment evaluates models by Pearson correlation, but this model can output either 0 or 1.

Evaluator	Accuracy	Macro-F_1
Global majority	0.564	0.361
Private majority	0.683	0.659
BERT-NSP	0.548	0.444
BERT-RUBER	0.541	0.488
Deep-AM	0.507	0.495
Deep-FM	0.543	0.533
Deep-AM-FM	0.541	0.531
FED	0.460	0.446
BERT-DCP	0.668	0.653
+ user token	**0.751**	**0.744**
+ profile	0.746	0.738
+ both	**0.751**	**0.744**

Table 4: Binary classification result of dialogue continuity prediction task on X dialogue dataset.

GPT-2[8] (small) for FED and Deep-FM. We trained models again for domain adaptation for FED and Deep-AM-FM, and additionally fine-tuned them for BERT-* using training data.[12] For our evaluator models, we trained BERT through the DCP task without the target user awareness (BERT-DCP) and with the personalization using user-specific token (+ user token), profile text (+ profile), or both of them (+ both). The hyperparameters of all models were as follows: learning rate as $3e - 5$, batch size as 64, and number of epochs as 5. We used AdamW (Kingma and Ba, 2015) as the optimizer and cross-entropy loss as the loss function. All model parameters trained on our dataset, including the annotator's conversation for the second experiment, are shared across all experiments.

4.3   Results

Table 4 lists the results of binary classification on the dialogue continuity prediction task in terms of accuracy and macro-F_1 to correct label bias. To compare the baseline model which does not output probabilities (Deep-AM-FM, FED), the model output is binarized using a threshold based on the whole user response ratio in the validation data. Unsurprisingly, BERT-DCP fine-tuned through DCP task performed better than the baselines. The evaluator can work with the DCP task by considering the interlocutor and get better results than **Private majority**. We also observed that using a unique speaker token for each interlocutor was a more effective way of taking interlocutors into account.

Table 5 lists the results of Pearson's r correla-

---

[12] We created their negative samples with the same amount of positive samples in the training data by randomly combining a dialogue context utterances and a reply.

Evaluator	Annotator 1		Annotator 2	
	Human	System	Human	System
BERT-NSP	0.477	**0.416**	−0.028	**0.337**
BERT-RUBER	0.285	0.243	−0.134	0.210
Deep-AM	0.564	0.293	0.015	0.242
Deep-FM	0.499	0.031	−0.011	−0.070
Deep-AM-FM	0.528	0.074	−0.010	−0.055
FED	0.210	0.051	0.040	−0.070
BERT-DCP	0.646	0.401	0.578	0.156
+ user token	0.720	0.364	**0.582**	0.072
+ profile	**0.754**	0.369	0.543	0.077
+ both	0.727	0.367	0.527	0.078

Table 5: Correlation with human judgment for responses by humans (Human) and dialogue systems (System).

Figure 3: Result of dialogue continuity prediction task per user group split according to training sample size.

tion between each evaluator's outputs (probabilities) and interlocutor scores. Our evaluators, BERT-DCP, have higher correlations with fluent human responses than baseline evaluators, and the improvement of performance by considering personality can be confirmed. This result confirms the usefulness of the DCP task for predicting interlocutor evaluations. In contrast, the BERT-NSP has the highest correlation in the system response, and all BERT-NSPs are worse than the performance in the human response. This may be because the DCP task is trained based on real conversations and is therefore vulnerable to non-fluent and inappropriate responses by the system. A similar tendency of lower correlation with human judgments for system responses than those for human responses has been reported for the other evaluation models on engagement (Ghazarian et al., 2020; Gao et al., 2020).

Because our interlocutor-aware evaluators correlate well with interlocutors' judgments of human responses, our method will be more useful as dialogue systems converse more naturally like humans. However, we still need to improve the evaluator so that it is capable of evaluating dialogue systems in the future.

4.4 Discussion

The performance of our interlocutor-aware evaluator will be affected by the size of the conversation logs given by the target interlocutor. For example, the performance could be poor for users who have a few conversations in the training data. To investigate the relationship between the training sample size for the target interlocutor and the performance of our models, we divide the test dataset into three user groups so that the training sample size for each group is as equal as possible. As a result, the aver-
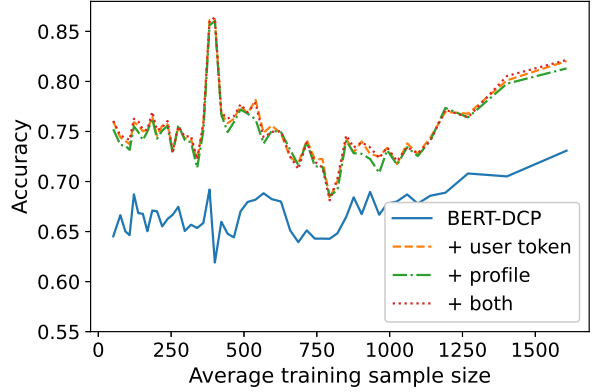
age sample size for each group was approximately 60,000, and the smallest group had an average of 51.1 samples. Table 3 shows the result on each user group in the test dataset. We confirmed that, with the exception of a peak around 400 samples, the accuracy changed only slightly below 1200 samples, improved above 1200 samples, and overall, the personalized models outperformed the BERT-DCP.

5 Conclusions

This study first explored the effect of interlocutor awareness on predicting interlocutor evaluations and then examined an automatic response evaluation method grounded in the perspective of the interlocutor. In the first experiment using the Hazumi dataset, we confirmed interlocutor score prediction requires personalization for interlocutor awareness as well as interlocutor scores. In the second experiment using conversations on X (formerly Twitter), we confirmed that dialogue continuity prediction is effective in training our interlocutor-aware automatic evaluator and the evaluator correlates with the actual interlocutor evaluations on human responses, while the improvement of the evaluation for the system responses is future work.

We plan to leverage recent response generation methods in long-term conversations (Xu et al., 2022b,c; Bae et al., 2022; Takasaki et al., 2023) to personalize our evaluator.

Limitations

Although this study illuminates the demand for evaluation from the perspective of the interlocutor, we only confirmed evaluation in terms of engagement. As existing studies on evaluation for open-domain dialogue systems are conducted in a variety of metrics such as understandability and informativeness, etc, (Finch et al., 2023), interlocutor-aware evaluation in the other evaluation metrics needs to be investigated.

To realize the study for a variety of metrics, a dataset with sufficient size of conversations and annotations is needed. In this study, we conducted experiments with two annotators to compare the automatic evaluators, but it is desirable to be annotated by a variety of people. Therefore, it is necessary to overcome the difficulties of the cost of constructing a dataset that includes conversations with multiple dialogue systems and annotations by the speakers, as well as the privacy issues related to dataset publication to reproduce experiments.

References

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. 2022. A personalized dialogue generator with implicit user persona detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 367–377, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alberto Díaz and Pablo Gervás. 2007. User-model based personalized summarization. *Information Processing & Management*, 43(6):1715–1734. Text Summarization.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don't forget your ABC's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071, Toronto, Canada. Association for Computational Linguistics.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *NIPS, modern machine learning and natural language processing workshop*, volume 2.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

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

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7789–7796.

Ao Guo, Atsumoto Ohashi, Ryu Hirai, Yuya Chiba, Yuiko Tsunomori, and Ryuichiro Higashinaka. 2021.

Influence of user personality on dialogue task performance: A case study using a rule-based dialogue system. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 263–270, Online. Association for Computational Linguistics.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

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

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations*.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. 2012. On recommending hashtags in twitter networks. In *Social Informatics*, pages 337–350, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Rajhans Mishra, Pradeep Kumar, and Bharat Bhasker. 2015. A web recommendation system considering sequential information. *Decision Support Systems*, 75:1–10.

Alexandros Papangelis, Nicole Chartier, Pankaj Rajan, Julia Hirschberg, and Dilek Hakkani-Tur. 2022. Understanding how people rate their conversations. In *Conversational AI for Natural Human-Centric Interaction, the 12th International Workshop on Spoken Dialogue System Technology*, IWSDS 2021,, pages 179–189, Singapore. Springer Singapore.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. volume 1, page 9.

Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2017. Modeling situations in neural chat bots. In *Proceedings of ACL 2017, Student Research Workshop*, pages 120–127, Vancouver, Canada. Association for Computational Linguistics.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems.

Meguru Takasaki, Naoki Yoshinaga, and Masashi Toyoda. 2023. Effective dialogue-context retriever for long-term open-domain conversation. In *The 13th International Workshop on Spoken Dialogue Systems Technology*, Los Angeles.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *AAAI Conference on Artificial Intelligence*, pages 722–729.

Yuma Tsuta, Naoki Yoshinaga, and Masashi Toyoda. 2020. uBLEU: Uncertainty-aware automatic evaluation method for open-domain dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 199–206, Online. Association for Computational Linguistics.

Guangxuan Xu, Ruibo Liu, Fabrice Harel-Canada, Nischal Reddy Chandra, and Nanyun Peng. 2022a. EnDex: Evaluation of dialogue engagingness at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4884–4893, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022b. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022c. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li. 2021. Deep amfm: Toolkit for automatic dialogue evaluation. In Luis Fernando D'Haro, Zoraida Callejas, and Satoshi Nakamura, editors, *Conversational Dialogue Systems for the Next Decade*, pages 53–69. Springer Singapore, Singapore.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.