General Paper Peer-Reviewed

Uncertainty-aware Automatic Evaluation Method for Open-domain Dialogue Systems

Yuma Tsuta[†], Naoki Yoshinaga[††] and Masashi Toyoda[††]

Because open-domain dialogues allow diverse responses, common reference-based metrics for text generation, such as BLEU, do not correlate well with human judgments unless we prepare an extensive reference set of high-quality responses for input utterances. In this study, we propose a fully automatic, uncertainty-aware evaluation method for open-domain dialogue systems, υBLEU. Our method first collects diverse reference responses from massive dialogue data, annotates their quality judgments by using a neural network trained on automatically collected training data, and then computes weighted BLEU using the automatically-retrieved and -rated reference responses. We also employ this method with an embedding-based metric, BERTScore, instead of the word-overlap-based metric, BLEU, to absorb surface variations of the reference responses. The experimental results on the meta-evaluation of our evaluation method for dialogue systems based on massive Twitter data confirmed that our method substantially improves correlations between BLEU (or BERTScore) and human judgments. We also confirmed that our method is effective when it is combined with a reference-free metric.

Key Words: *Open-domain Dialogue System, Evaluation Method*

1 Introduction

Interest in intelligent dialogue agents (i.e., Apple Siri, Amazon Alexa, and Google Assistant) has increased. The key to achieving higher user engagement with the dialogue agents is to support open-domain non-task-oriented dialogues to return a meaningful response for any user input. Along with an increase of online conversational data on social media platforms such as Twitter and Reddit and the advent of deep neural networks, researchers study data-driven approaches to open-domain dialogue systems (Ritter et al. 2010; Vinyals and Le 2015; Shang

[†] Graduate School of Information Science and Technology, The University of Tokyo
[††] Institute of Industrial Science, The University of Tokyo
Part of this study was presented at proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop.

et al. 2015).

The major challenge in developing open-domain dialogue systems is that existing evaluation metrics for text generation tasks, such as BLEU (Papineni et al. 2002), correlate poorly with human judgment in evaluating responses generated by dialogue systems (Liu et al. 2016). In open-domain dialogues, even though responses with various contents and styles are acceptable (Sato et al. 2017), only a few responses, or often only one, are available as reference responses in evaluation datasets because the datasets are made from actual online conversations. Therefore, it is hard for these reference-based metrics to consider uncertain responses without writing additional reference responses by hand (Li et al. 2017) (§ 2).

To address the weak correlation between existing reference-based metrics and human evaluation, some researchers resort to reference-free metrics (Tao et al. 2018; Mehri and Eskenazi 2020a, 2020b), while other researchers augment reference-based metrics with embedding-based soft matching (instead of word overlaps used in BLEU) (Zhang et al. 2020) or more directly with multiple automatically-extended reference responses (Sordoni et al. 2015). However, the task of evaluating responses without references is inherently as difficult as generating responses themselves and requires extra supervision to obtain a high correlation with human judgments (Lowe et al. 2017; Sai et al. 2020). Although embedding-based soft matching can partially address the diversity of outputs in open-domain dialogues, it is useless to evaluate responses that have different semantic content with available references. Galley et al. (2015) proposed ΔBLEU (§ 3.2), which integrates human judgments on automatically-extended reference responses with diverse quality in BLEU computation. Although this method shows promising results, it requires human judgments to score automatically-extended references. Therefore, it cannot effectively evaluate open-domain dialogue systems in a wide range of domains of interest.

In this study, to obtain robust reference-based metrics for open-dialogue systems, we propose an automatic, uncertainty-aware evaluation metric, υBLEU, which removes the human intervention in ΔBLEU. The proposed metric exploits reference responses that are retrieved from massive dialogue logs and are automatically rated by a neural network (§ 4). We first retrieve diverse response candidates according to the similarity of utterances to which the responses were directed. Next, we evaluate the appropriateness of each retrieved response using a neural network. We train the neural network model using data automatically generated from an utterance with multiple responses. Finally, we use the retrieved responses and their automatic scores to evaluate the target response with a weighted BLEU following Galley et al. (2015). In addition, we propose two extensions: first, an improvement of the overall method using BERT (Devlin et al. 2019) and its derivative, the embedding-based metric BERTScore (Zhang et al. 2020) (§ 3.3); second, an inte-

gration with the evaluation metric RUBER (Tao et al. 2018) (§ 3.4), which combines reference-free and reference-based metrics, to demonstrate the usefulness of our method when combined with reference-free metrics.

Using our method, we experimentally evaluated responses generated by dialogue systems such as a retrieval-based method (Liu et al. 2016), a generation-based method (Serban et al. 2017), and human as an ideal system[1] using Twitter dialogues (§ 5). Our method is comparable to ΔBLEU in terms of its correlation with human judgment, and when it is integrated into RUBER (Tao et al. 2018), it substantially improves the correlation (§ 6).

Our contributions are as follows.

- We developed a fully-automatic uncertainty-aware evaluation method for open-domain dialogue systems. Our method automates the human ratings required in ΔBLEU while maintaining the performance.
- We demonstrated that even a recent embedding-based reference-based metric, BERTScore, can benefit from the extended reference responses that are retrieved and then rated by our method.
- We confirmed that combining vBLEU with the unreferenced scorer of RUBER (Tao et al. 2018) improves RUBER's performance by considering the diversity of outputs in open-domain dialogues.

2 Related Work

Here, we introduce recent studies on evaluating open-domain dialogue systems. We focus on model-agnostic methods that can evaluate a system response for a given utterance.[2]

2.1 Reference-based metrics

For the evaluation of dialogue systems, researchers have first adopted evaluation metrics that have been designed for other text-generation tasks. Specifically, BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005) for machine translation and ROUGE (Lin 2004) for automatic summarization are used to evaluate open-domain dialogue systems. However, these word-overlap-based metrics are hard to capture semantic similarities between words and have been reported to correlate poorly with human judgments (Liu et al. 2016).

[1] To use human responses, we collect utterances with multiple responses as a test dataset.

[2] Perplexity is sometimes used to evaluate dialogue systems (Hashimoto et al. 2019). However, it is only applicable to generation-based dialogue systems; therefore, we do not discuss it here, like Liu et al. (2016).

Embedding-based metrics such as BERTScore (Zhang et al. 2020) remedy the shortcomings of word-overlap-based metrics by exploiting word embeddings that capture word meanings with continuous vector representations. Unfortunately, even using embedding-based similarities, reference-based metrics correlate poorly with human judgments in evaluating dialogue systems (Liu et al. 2016; Ji et al. 2022). This is because reference-based metrics are computed with only single reference responses taken from actual conversations, whereas acceptable responses in open-domain dialogues can be diverse (Sato et al. 2017).

To directly capture the diversity of outputs in open-domain dialogues, Sordoni et al. (2015) attempted to collect multiple reference responses from dialogue logs for each test utterance-response pair. Because automatically-augmented reference responses are noisy, Galley et al. (2015) improved this method by manually rating the augmented reference responses and used the ratings to perform discriminative BLEU evaluation, as detailed later in § 3.2. Gupta et al. (2019) created multiple reference responses by hand for the DailyDialog dataset (Li et al. 2017) to confirm the effect of multiple reference responses for evaluating dialogue systems. Although these studies confirmed that human-curated reference responses improve the correlation with human judgments, it is costly to create such evaluation datasets for various domains.

In this study, motivated by the moderate correlation with human judgments obtained using multiple human-curated references, we designed a fully automatic uncertainty-aware evaluation method for dialogue systems. We confirmed that the embedding-based metric, BERTScore, can take advantage of multiple references that are retrieved and rated by our method.

2.2 Reference-free metrics

Because reference-based metrics with single reference responses do not correlate well with human judgments, various researchers have explored methods to evaluate dialogue systems without relying on reference responses (Tao et al. 2018; Ghazarian et al. 2019; Mehri and Eskenazi 2020a, 2020b; Sinha et al. 2020; Sai et al. 2020). The referenced-free methods learn to model the relevance between an utterance and a response in a supervised or unsupervised manner.

RUBER (Tao et al. 2018) is an automatic evaluation method that combines two approaches: its referenced scorer evaluates the similarity between a reference and a generated response using the cosine similarity of their vector representations, whereas its unreferenced scorer, trained by negative sampling, evaluates the relevance between an input utterance and a generated response. Ghazarian et al. (2019) demonstrated that contextualized word embeddings (Devlin et al. 2019) improve the unreferenced scorer but not the referenced scorer in RUBER. The referenced scorer is similar to BLEU in that they both are referenced-based evaluation metrics.

Some researchers exploited the language recognition ability of large-scale pre-trained models in the evaluation of open-domain dialogue systems. USL-H (Phy et al. 2020) and USR (Mehri and Eskenazi 2020b) evaluate the fluency of generated responses from the probabilities of token output by transforming the masked language model task of BERT and the appropriateness of system responses to the context by transforming the next sequence prediction task of BERT. Deep AM-FM (Zhang et al. 2021) uses the perplexity of the text generation-based pre-trained language model as the evaluation score of the generated responses. FED (Mehri and Eskenazi 2020a) calculates the perplexity of a given follow-up utterance such as "Wow! Very interesting" to the generated responses as an evaluation score.

The above reference-free metrics are orthogonal efforts to our uncertainty-aware method based on automatically-retrieved and rated reference responses. We evaluated RUBER using our method as a referenced scorer to show the utility to combine reference-free metrics with reference-based metrics with multiple references rated by a neural network (§ 5.5).

3 Preliminaries

Here, we review ΔBLEU (Galley et al. 2015), a human-aided evaluation method for text generation tasks with uncertain outputs, after explaining the underlying metric, BLEU (Papineni et al. 2002). We then explain the embedding-based metric BERTScore (Zhang et al. 2020) and RUBER (Tao et al. 2018) that utilizes reference-free metric because each of metric is combined with our method later in § 4.3 and § 5.5.

3.1 BLEU

BLEU (Papineni et al. 2002) calculates an evaluation score based on the number of occurrences of n-gram tokens that appear in both reference response r and generated response h. Specifically, the score is calculated from a modified n-gram precision p_n and a brevity penalty BP

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_n \frac{1}{N} \log p_n\right), \tag{1}$$

$$\text{BP} = \begin{cases} 1 & \text{if } \eta > \rho \\ e^{(1-\rho/\eta)} & \text{otherwise} \end{cases}, \tag{2}$$

$$p_n = \frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{\#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(h_i)} \#_g(h_i)}. \tag{3}$$

Here, ρ and η are the average lengths of the reference and generated responses, respectively; n and N are the n-gram length and its maximum, respectively; h_i and $\{r_{i,j}\}$ are the generated response and the j th reference response[3] for the i th utterance, respectively; $\#_g(u)$ is the number of occurrences of n-gram token g in sentence u; $\#_g(u, v)$ is defined as $\min\{\#_g(u), \#_g(v)\}$.

3.2 ΔBLEU: Discriminative BLEU

ΔBLEU (Galley et al. 2015) is a human-aided evaluation method for text generation tasks with uncertain outputs, such as response generation in open-domain dialogues. To augment the reference responses for each test example (an utterance-response pair), following the work by Sordoni et al. (2015), ΔBLEU first retrieves utterance-response pairs similar to the given pair from conversation logs in SNS such as Twitter. They compute the similarity between the test sample and each utterance-response pair in the retrieval pool by multiplying the similarity between the utterances and the similarity between the responses, each of which is computed by BM25 (Robertson et al. 1995). Next, the responses of the top 15 similar utterance-response pairs and the utterance itself (as a parrot return) are combined with the original response to form an extended set of reference responses. Each of the extended references is then rated by humans, in terms of its appropriateness as a response to the given utterance. Finally, ΔBLEU calculates p_n (Eq. 3) with the extended reference $r_{i,j}$ and its manual quality judgment $w_{i,j}$ for the input utterance i:

$$p_n = \frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_{j: g \in r_{i,j}} \{w_{i,j} \cdot \#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{w_{i,j} \cdot \#_g(h_i)\}}. \tag{4}$$

In this way, ΔBLEU weights the number of occurrences of n-gram g in Eq. 3 with manual quality judgment $w_{i,j}$.

The problem with ΔBLEU is the cost of human judgments. Although we want to evaluate open-domain dialogue systems in various domains, the annotation cost prevents effective evaluation.

3.3 BERTScore

BERTScore (Zhang et al. 2020) is a reference-based evaluation method for text generation tasks that performs evaluation calculations by soft matching of contextual word embeddings of BERT (Devlin et al. 2019). BERTScore uses a pre-trained BERT and this allows domain- and task-independent evaluation. Given a reference sentence r with the tokens of $r_1, \ldots, r_k$ and a

[3] BLEU allows multiple reference responses for a single utterance (cf. § 2.2.1 in Papineni et al. (2002)).

candidate sentence h with the tokens of $h_1, \ldots, h_l$, BERTScore first computes BERT embeddings of the tokens ($\langle \mathbf{r}_1, \ldots, \mathbf{r}_k \rangle$ and $\langle \mathbf{h}_1, \ldots, \mathbf{h}_l \rangle$) by inputting these sentences independently into BERT. Thereafter, BERTScore obtains the evaluation scores according to the following formula:

$$P_{\mathrm{BERT}} = \frac{1}{|h|} \sum_{h_j \in h} \max_{r_i \in r} \mathbf{r}_i^\top \mathbf{h}_j, \tag{5}$$

$$R_{\mathrm{BERT}} = \frac{1}{|r|} \sum_{r_i \in r} \max_{h_j \in h} \mathbf{r}_i^\top \mathbf{h}_j, \tag{6}$$

$$F_{\mathrm{BERT}} = \frac{2 \cdot P_{\mathrm{BERT}} \cdot R_{\mathrm{BERT}}}{P_{\mathrm{BERT}} + R_{\mathrm{BERT}}} \ . \tag{7}$$

If multiple reference responses are available, BERTScore adopts the highest evaluation score after calculating the scores using each reference response. The use of multiple references is mentioned in the evaluation task of image caption generation in the original paper (cf. § 4 in Zhang et al. (2020)).

3.4 RUBER

RUBER (Tao et al. 2018) is an automatic evaluation method for open-domain dialogue systems that combines two types of scorers: an unreferenced scorer, which calculates a relevance score between the given utterance and the system response, and a referenced scorer, which evaluates similarity between the system responses and the reference response.

The unreferenced scorer is implemented with a neural network model that takes a generated response h and the query utterance q, and returns a score s_U by the following process: first, the generated response h and the quey utterance q are independently vectorized into $\mathbf{V}_h$ and $\mathbf{V}_q$ using bi-directional GRU (Bi-GRU) (Eq. 8). Thereafter, a quadratic feature m is obtained from $\mathbf{V}_h$ and $\mathbf{V}_q$ using a learnable parameter matrix $\mathbf{M}$ (Eq. 9). Finally, multi-layer perceptron (MLP) with a sigmoid function is fed to the joint vector $[\mathbf{V}_h; m; \mathbf{V}_q]$ and outputs a score s_U (Eq. 10).

$$\mathbf{V}_h = \text{Bi-GRU}(h), \ \mathbf{V}_q = \text{Bi-GRU}(q), \tag{8}$$

$$m = \mathbf{V}_h \cdot \mathbf{M} \cdot \mathbf{V}_q, \tag{9}$$

$$s_U = \sigma(\text{MLP}([\mathbf{V}_h; m; \mathbf{V}_q])). \tag{10}$$

The model is optimized with margin loss J using negative sampling with margin Δ.

$$J = \max(0, \Delta - s_U(q, r) + s_U(q, r^-)),$$

where Δ is a hyper-parameter and r^- denotes a randomly selected response.

The referenced scorer, s_R, is an evaluation metric derived from the cosine similarity of sentence

embeddings between generated response h and reference response r. In this metric, a sentence embedding $\mathbf{v}_X$ of sentence X is obtained from the concatenation of the maximum and minimum pooling for the word embeddings $\{\mathbf{w}_0, \ldots, \mathbf{w}_n\}$ of X.

$$\mathbf{v}_{\max}[i] = \max(\mathbf{w}_0[i], \ldots, \mathbf{w}_n[i]),$$
$$\mathbf{v}_{\min}[i] = \min(\mathbf{w}_0[i], \ldots, \mathbf{w}_n[i]),$$
$$\mathbf{v}_X = [\mathbf{v}_{\max}; \mathbf{v}_{\min}],$$

where $[\bullet]$ indexes a dimension of an embedding.

RUBER normalizes scores $s_U(q, h)$ and $s_R(r, h)$ (for a generated response h to the query utterance q and the reference response r) to the range of $[0, 1]$ before summarizing them to obtain summary score $s(q, r, h)$. The normalization for each score is given by

$$\tilde{s}_* = \frac{s_* - \max(S)}{\min(S) - \max(S)}$$

where s_* means either s_U or s_R, and S means the set of scores for each metric in the evaluation sample set. Summary score $s(q, r, h)$ for the normalized scores ($\tilde{s}_U(q, h)$ and $\tilde{s}_R(r, h)$) are obtained using maximum or minimum pooling, or arithmetic or geometric mean. In this study, the arithmetic mean was employed because the result on the original paper is the most stable.

4 Uncertainty-aware Evaluation Method for Dialogue Systems

Here, we describe our approach to the problem of ΔBLEU described in § 3.2. To remove the cost of human judgments of extended references, we propose using a neural network trained on automatically collected training data to rate each of the retrieved responses (Figure 1, § 4.2). In addition, to diversify the extended reference responses in terms of content and style, we propose a relaxed response retrieval approach using continuous vector representations of utterances only (§ 4.1). Additionally, we introduce an extension of our method that uses a pre-trained language model, which is effective for various classification tasks (§ 4.3). In this part, we aim to improve the entire model by modifying the proposed method and ΔBLEU's weighted BLEU to a method using pre-trained language models.

4.1 Retrieving diverse reference responses

Given an utterance-response pair (test example), ΔBLEU expands the original reference response by retrieving utterance-response pairs, in which both the utterance and response are similar to the test example, from massive dialogue logs (here, Twitter). Because using the similarity

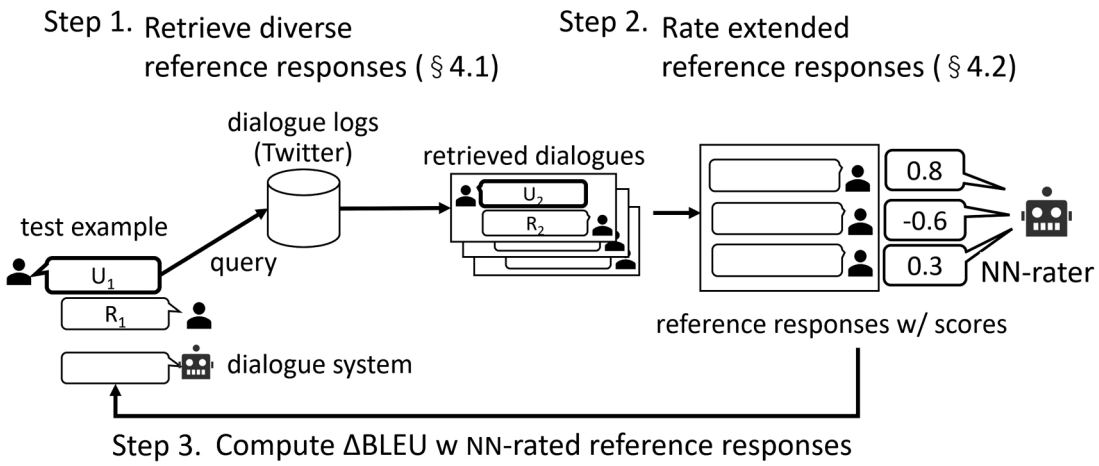

Figure 1 Overview of υBLEU: retrieving diverse reference responses from dialogue logs using utterance U_1 only (§ 4.1) to augment the reference response R_1 in each test example, validating their quality by neural network (NN)-rater (§ 4.2) and evaluating the responses generated by the dialogue system using ΔBLEU.

between responses prevents us from retrieving diverse responses in terms of content, we propose considering only the similarity between the utterances. In addition, we use an embedding-based similarity instead of BM25 to flexibly retrieve semantically-similar responses with synonymous expressions (style variants).

We compute the similarity of utterances using the cosine similarity between utterance vectors obtained from the average of pre-trained embeddings of the words in the utterances. In addition to the retrieved responses, we add the utterance (as a parrot return) to the reference responses as in ΔBLEU.

4.2 Rating extended reference responses

ΔBLEU manually evaluates the appropriateness of the extended reference responses for the utterance. To remove this human intervention, we propose rating each reference response using a neural network that outputs a probability for that response as a response to the given utterance.

Specifically, our neural network (NN)-rater takes two utterance-response pairs as inputs: a given pair of utterance U_1 and reference response R_1 (test example), and a retrieved pair of utterance U_2 and response R_2 (Figure 2). The NN-rater is trained to output the probability that the retrieved response R_2 for U_2 can be a response to given utterance U_1 with response R_1. This probability is then used as a quality judgment after normalization to the interval $[-1, 1]$ as in ΔBLEU.

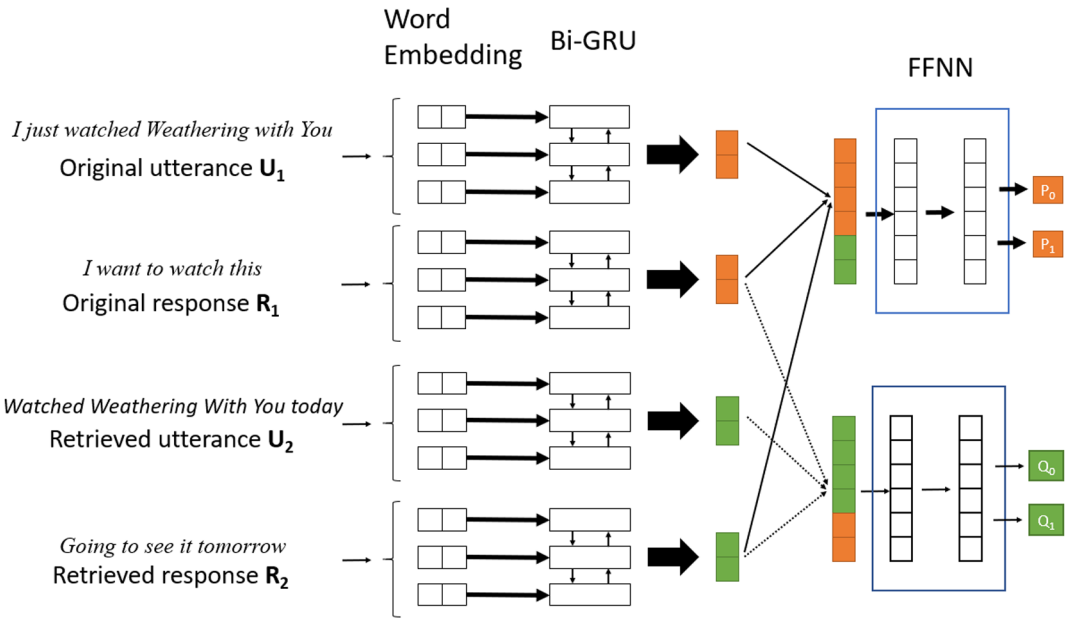

Figure 2 Overview of NN-rater (§ 4.2): calculating the quality of response R_2 to utterance U_1, inputting a triplet of a pair of an utterance U_1 and a response R_1 of some dialogue and a response R_2 of other dialogue into neural network model of Bi-GRU and FFNN. In the dialogue on estimation and the dialogue of negative label in training data, U_1 and U_2 are different as shown in this figure; however, the U_1 and U_2 of the dialogue of positive label in training data are the same, unlike this figure.

The key issue here is how to prepare the training data for the NN-rater. We used utterances with multiple responses in dialogue data (here, Twitter) as positive examples; we randomly sampled two utterance-response pairs for negative examples.

We then trained the NN-rater in Figure 2 from the collected training data. Because the utterances in the two utterance-response pairs in a positive example are identical while those in a negative example are independent, we do not feed both utterances to the NN-rater. This input design prevents overfitting.

We constructed NN-rater as a simple model that takes three sentences, concatenates them after vectorization, and outputs the probability of a label. Specifically, given a test example of utterance U_1 and response R_1 and a retrieved utterance-response pair of U_2 and R_2, we give two triplets, $\langle U_1, R_1, R_2 \rangle$ and $\langle U_2, R_2, R_1 \rangle$, as inputs to the NN-rater. Next, we made two vectors by concatenating triplet vectors returned from the bi-directional gated recurrent unit (Bi-GRU) (Cho et al. 2014) as the last hidden state for the utterance and the two responses. We concatenated

forward and backward hidden states (h_f, h_b) in the Bi-GRU to represent an utterance/response vector as $v = [h_f, h_b]$. We then feed each triplet vector to a feed-forward neural network (FFNN) with softmax function to obtain a pair of probabilities that R_2 can be a response to U_1 or not (similarity, another pair of probabilities that R_1 can be a response to U_2 or not). The maximum of these two probabilities is used as the qualitative judgment of the response R_2 (or R_1) and multiplied by -1 if classified as negative to normalize into $[-1, 1]$.

4.3 υBERTScore: Extending υBLEU with BERT

Recently, large-scale pre-trained models such as BERT (Devlin et al. 2019) are used in various NLP tasks as-is or as foundation models and outperform existing methods. Hence, as an extension of υBLEU above, we propose υBERTScore, which employs pre-trained models at each step. We used BERT to capture meanings of utterances accurately in retrieving and rating extended references and use BERTScore (Zhang et al. 2020) instead of BLEU to perform soft matching between generated responses and the extended references. We explain how to extend υBLEU using BERT below.

Dialogue retrieval with contextualized embeddings computed by BERT First, instead of using a pre-trained word embedding to collect similar dialogues to a query (§ 4.1), we adopted the output of each token from BERT as a word embedding. Because BERT is trained as a language model, outputs of each token from BERT are contextualized to input sentence compared to a pre-trained word embedding such as GloVe (Pennington et al. 2014). Therefore, retrieval with the output from BERT may recognize phrases and style, as well as word meanings.

BERT as NN-rater Second, in the validation for extended reference responses in υBLEU, we used BERT as an implementation of NN-rater instead of Bi-GRU and FFNN. As we finetune BERT, which is pre-trained as a language model in a large corpus, as a validator, we can expect high classification performance based on the high language recognition ability.

BERTScore with weighted reference responses Finally, we utilized BERTScore, which performs soft matching between system outputs and reference outputs using contextualized word embeddings, instead of BLEU, which is based on word overlaps. To weight each extended reference response by its appropriateness score in the BERTScore, we changed the calculation method (Eq. 5, 6) to the following method with the validity score w to reference sentence r.

$$P_{\mathrm{BERT}} = \frac{1}{|h|} \sum_{h_j \in h} \max_{r_i \in r} w \mathbf{r}_i^\top \mathbf{h}_j, \tag{11}$$

$$R_{\mathrm{BERT}} = \frac{1}{|r|} \sum_{r_i \in r} \max_{h_j \in h} w \mathbf{r}_i^\top \mathbf{h}_j. \tag{12}$$

5 Experimental settings

Here, we describe how to evaluate our methods for evaluating open-domain dialogue systems. Using utterances from Twitter (§ 5.1), responses written by humans, and responses obtained by dialogue systems (§ 5.2), we evaluated our methods in terms of their correlation with human judgment (Table 1, § 5.4–5.5).

5.1 Twitter dialogue datasets

We developed a large Japanese dialogue dataset from the Twitter archive that we have collected since March 2011. This archive consists of regularly collected posts by tracked users through the Twitter API.[4] We initially selected approximately 30 well-known Japanese users to be tracked, and by adding users mentioned or retweeted by these tracked users to be tracked, the number of users has expanded to approximately 2.5 million. To create a dataset from this archive, we extracted only Japanese-language posts, pre-processed the posts to denoise them, and then obtained conversation logs.

To extract Japanese language posts, we used the language labels provided by the API and ldig,[5] a language classification model specifically designed for Twitter, to increase reliability. ldig can classify the language of posts on Twitter with more than 99% accuracy for 17 languages. Posts classified as Japanese by both models were used from the archive.

To denoise posts, non-linguistic information such as URLs and emojis were removed from the posts. We also removed or omitted SNS-specific text, such as user names (e.g., @user_name) and repeated letters or words. Finally, we removed posts that appear to be from bots or have little content, such as posts that contain less than 5 characters, do not contain more than 60% Japanese, or are duplicates on the same day.

Finally, we constructed a conversational dataset from the denoised posts. Specifically, we

Models	Similarity on Retreival	Validation	Accumulation
ΔBLEU	Dialogue matching with BM25	HUMAN	Weighted-BLEU
υBLEU	Utterance matching with GloVe	NN-rater (Bi-GRU & FFNN)	Weighted-BLEU
υBERTScore	Utterance matching with BERT	NN-rater (BERT)	Weighted-BERTScore

Table 1 Summary of each multi-reference-based evaluation method compared in the experiments.

[4] https://developer.twitter.com/en/docs/twitter-api
[5] https://github.com/shuyo/ldig

Models (components)	Training	Validation	Test
VHERD C-BM25 (training data used as retrieval pool)	2.4M (2018)	10K (2018) n/a	100 (2019)
NN-rater, RUBER	5.6M (2017)	10K (2017)	n/a
υBLEU (extended-references retrieval, training GloVe, and DAPT for BERT)	Approximately 16M (2017)		

Table 2 Statistics of the dialogue data used to run each model (component). The numbers in the parentheses denote the year.

considered posts that were neither retweets nor references to other posts as utterances and posts that referred to those posts as responses.

We used this dataset for training and testing dialogue systems and for training the NN-rater that evaluates the quality of retrieved responses. In these experiments, to simulate evaluating dialogue systems trained with dialogue data that are unseen by evaluation methods, we used dialogue data posted during 2017 for training and running the NN-rater, and dialogue data posted during 2018 for training and during 2019 for testing the dialogue systems, as summarized in Table 2.

5.2 Target responses for evaluation

Following Liu et al. (2016) and Lowe et al. (2017), we adopted three methods to obtain responses for each utterance in the test set to evaluate various types of dialogue systems: a retrieval-based method C-TFIDF (Liu et al. 2016) with BM25 as the similarity function (C-BM25), a generation-based method VHRED (Serban et al. 2017), and HUMAN responses, which are the actual responses except for the reference response, as an ideal response generated by a dialogue system.

Following Ritter et al. (2010) and Higashinaka et al. (2011), to extract various dialogues from Twitter as training data for the above methods, we recursively follow replies from each non-reply post to obtain a dialogue between two users that consists of at least three posts. We then randomly selected pairs of the first utterances and their replies in the obtained dialogues as our dialogue data: 2.4M pairs for training the VHRED and for retrieving responses in C-BM25, 10K pairs as validation data for the VHRED, and 100 pairs as test data.[6] These dialogues were tokenized with SentencePiece (Kudo and Richardson 2018)[7] for VHRED and with MeCab 0.996

[6] To obtain HUMAN responses for evaluation, we only used dialogues whose first utterances had more than one response.

[7] The model was trained with a 16,000 vocabulary size on the training dataset.

(ipadic 2.7.0)[8] for C-BM25 to retrieve responses based on words that are less ambiguous than subwords.

Finally, six Japanese native speakers in our research group, who are not related to this research project, evaluated the randomly shuffled 300 target responses for the 100 test examples in terms of the appropriateness as a response to a given utterance. We used a 5-point Likert-type scale, with 1 denoting inappropriate or unrecognizable and 5 denoting very appropriate or seeming to be an actual response.

5.3 Response retrieval and scoring

Following Galley et al. (2015), for each test example, the 15 most similar utterance-response pairs were retrieved to augment the reference response in addition to the utterance itself (as a parrot return) to apply ΔBLEU, υBLEU, and υBERTScore. We retrieved utterance-response pairs from approximately 16M utterance-response pairs of our dialogue data (Table 2). These dialogue data were tokenized with MeCab for response retrieval; we then trained GloVe embeddings (Pennington et al. 2014) to compute utterance or response vectors (§ 4.1) and performed domain-adaptive pre-training (DAPT) (Gururangan et al. 2020) of a pre-trained BERT[9] on this dialogue data. We used the obtained domain-adapted pre-trained BERT for υBERTScore to retrieve extended reference responses and to build a BERT-based NN-rater as explained in the section below.

5.4 NN-rater to evaluate reference responses

To train the NN-rater for evaluating the extended references (§ 4.2), we randomly extracted 5.6M and 10K utterance-response pairs for training and validation data, respectively. The number of positive and negative examples was set equal in both data. Before these examples were fed to the NN-rater in υBLEU, they are tokenized with SentencePiece.

For the NN-rater in υBLEU, we used a 512-dimensional embedding layer, one Bi-GRU layer with 512-dimensional hidden units, five layers for the FFNN with 1024-dimensional hidden units, and a ReLU as the activation function. We used the Adam optimizer (Kingma and Ba 2015) with an initial learning rate of 0.001 and calculated the loss by the cross entropy. We trained the NN-rater with a batch size of 1,000 and up to 15 epochs. The model with parameters that achieved the minimum loss on the validation data was used for evaluating the test data.

For the NN-rater in υBERTScore, we used the same architecture as the original BERT base model (Devlin et al. 2019); 12 layers, 768 dimensions of hidden states, and 12 attention heads. This

[8] https://taku910.github.io/mecab/
[9] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

model was pre-trained on Japanese Wikipedia as of September 1, 2019, and the text file used for training is 2.6GB in size, consisting of approximately 17M sentences. See the link for the other details. To obtain a domain-adapted pre-trained BERT, we trained the pre-trained BERT again using the training dataset for NN-rater. In addition, this domain-adapted BERT was trained with the same settings as above to obtain NN-rater.

We then evaluated the appropriateness of each retrieved reference response by humans for ΔBLEU and by the NN-raters for υBLEU and υBERTScore in terms of appropriateness as a response to a given utterance. We asked four of the six Japanese native speakers to evaluate the quality of each retrieved reference response.[10]

5.5 Compared evaluation methods for open-domain dialogue systems

To observe the impact of two modifications on ΔBLEU, namely more diverse reference retrieval (§ 4.1) and automatic reference quality judgment (§ 4.2), we first compared BLEU with various reference retrieval methods. A BERT-based reference retrieval method that uses contextualized embedding from BERT to retrieve dialogues based only on the cosine similarity of utterances is also compared. Thereafter, we compared BLEU and BERTScore with only one reference, ΔBLEU, and υBLEU, and υBERTScore. We then compared our proposed methods, υBLEU and υBERTScore, with RUBER, which uses a reference-free metric, and examined the performance of RUBER when its referenced scorer was replaced with our proposed methods.

Specifically, we applied each evaluation method to the 300 responses (§ 5.2). ΔBLEU, υBLEU, and υBERTScore used the extended references of each method in the evaluation. BLEU used the original (single) references or the extended references. The referenced scorer in RUBER used the original (single) references with GloVe embeddings,[11] as described in § 5.3

We evaluated the performance of the evaluation methods in terms of their correlation to human judgments on the 300 responses. To calculate the correlation, we used Spearman's ρ and Pearson's r. In addition to the correlations with the averaged human judgments, we computed the maximum and minimum correlation with human judgments given by each annotator to understand the stability of the evaluation. All evaluation methods using the modified n-gram precision were calculated with $n \leq 2$ (BLEU-2), following (Galley et al. 2015).

[10] We asked only part of them because of the large annotation size and hence the high cost.

[11] We changed a word embedding from word2vec (Mikolov et al. 2013), which was used in the experiments in the original paper (Tao et al. 2018). We confirmed that this change improves the performance of the referenced scorer in our experiment.

6 Results

Here, we present the results of the experiments, from the modifications to ΔBLEU to the combination with RUBER and other analyses. We first present a comparison of the reference retrieval methods in terms of the performance of BLEU using multiple references extended by each method. Subsequently, we present a comparison of reference-based metrics including ΔBLEU, vBLEU, and vBERTScore. We then demonstrate the performance of the reference-based metrics combined with RUBER('s unreferenced scorer, which is a reference-free metric). We finally analyze the performance of the NN-rater in terms of agreement with the human evaluation and qualitative analysis of each method.

Table 3 lists the correlations between the human judgments and BLEU scores for each reference retrieval method. We can observe that cosine similarity on embeddings correlates with human judgments more than BM25. When cosine similarity on embeddings is used, using only the utterances as keys outperformed using both utterance and response. When we compare the single and multiple reference-based metrics, for Pearson's r, only the proposed retrieval method, which uses an embedding-based similarity to utterances, showed a higher minimum correlation than BLEU did with only one reference. These results confirmed that the proposed retrieval methods were effective for extending the reference responses. Additionally, a comparison of embedding types confirmed that the BERT-based extension was effective for dialogue retrieval.

Table 4 compares the evaluation methods based on reference-based metrics. First, we established that the single-reference metrics (BERTScore and the referenced scorer in RUBER) did not correlate well with human judgments, as did single-reference BLEU in Table 3. Conversely, multi-

Methods to retrieve extended reference responses		Spearman's ρ			Pearson's r		
Key	Similarity funciton	max	min	avg	max	min	avg
(Only a single reference response)		.186	.091	.153	.276	.190	.242
Utterance & Response	BM25	.257	.138	.232	.298	.173	.261
Utterance only	BM25	.265	.136	.236	.296	.178	,267
Utterance & Response	Cosine similarity on GloVe vectors	.280	.148	.265	.322	.177	.301
Utterance only	Cosine similarity on GloVe vectors	.333	.181	.297	.366	.209	.338
Utterance only	Cosine similarity on BERT vectors	**.344**	**.192**	**.315**	**.387**	**.228**	**.360**

Table 3 Correlation between human judgment and BLEU with a single or extended reference responses; the extended reference responses are obtained by chaining the key (for utterance-response pairs) to compute similarities and similarity functions.

Metric	Spearman's ρ			Pearson's r		
	max	min	avg	max	min	avg
Single reference metrics						
BLEU	.186	.091	.153	.276	.190	.242
RUBER (Referenced Scorer)	.188	.071	.096	.075	.016	.060
BERTScore	.256	.155	.257	.342	.230	.331
Multi reference metrics						
ΔBLEU	.366	**.300**	.359	.360	.294	.353
υBLEU	.330	.281	.334	.394	**.332**	.371
υBERTScore	**.378**	.280	**.403**	**.426**	.316	**.431**
Human	.773	.628	—	.778	.607	—

Table 4 Correlation between reference-based method and human judgment; human refers to the inter-rater correlations. RUBER (Referenced Scorer) and BERTScore use the original single reference responses, whereas ΔBLEU, υBLEU, and υBERTScore use the extended reference responses with automatic ratings by each method.

reference metrics ΔBLEU, υBLEU, and υBERTScore correlated more than those. The comparison between υBLEU and BLEU in Table 3 revealed that our NN-rater improved the minimum correlation with human judgment. Here, υBLEU was comparable to ΔBLEU, which implies that our method can successfully automate ΔBLEU, a human-aided, uncertainty-aware evaluation method. We observed the highest minimum correlations with human judgments by ΔBLEU on Spearman's ρ and υBLEU on Pearson's r. This suggests that these metrics are stable evaluations that are easy to agree with various people. Conversely, the highest correlation with averaged human judgments by υBERTScore indicated that this metric is the best standard evaluation method. The comparison between BERTScore and υBERTScore also demonstrated that our reference augmentation is effective for methods other than BLEU.

Finally, Table 5 presents the results of a reference-free metric (the unreferenced scorer in RUBER) and combination methods of reference-based and reference-free metrics as an automatic evaluation method. The original RUBER combination performed worse than the (reference-free) unreferenced scorer owing to the poor performance of the (reference-based) referenced scorer (c.f. Table 4). Although this result is not consistent with the results reported in the original paper (Tao et al. 2018), Ghazarian et al. (2019) have failed to reproduce their results. By replacing the referenced scorer of RUBER with our proposed reference-based automatic methods, we obtained better overall correlations and the best performance with υBERTScore. This confirms that our

Metric	Spearman's ρ			Pearson's r		
	max	min	avg	max	min	avg
RUBER (Unreferenced Scorer)	.342	.225	.332	.336	.217	.325
RUBER (Unreferenced & Referecened Scorer)	.339	.206	.320	.325	.193	.307
RUBER (Unreferenced Scorer) & υBLEU	.435	.323	.440	.450	**.338**	.456
RUBER (Unreferenced Scorer) & υBERTScore	**.453**	**.327**	**.483**	**.474**	.336	**.473**
Human	.773	.628	—	.778	.607	—

Table 5 Correlation of human judgment with reference-free method or combination with reference-based method; human refers to the inter-rater correlations. υBLEU and υBERTScore use the extended reference responses with automatic ratings by each method.

uncertainty-aware reference-based metrics are also effective when combined with reference-free metrics.

Analysis: Evaluation of NN-rater In the proposed methods (υBLEU and υBERTScore), the reference response augmentation method and validation method to the retrieved responses were changed from ΔBLEU and BERTScore. The improvement of the reference response augmentation method was evaluated to be effective in increasing the correlation with human judgments in Table 3. However, the extent to which the validator (NN-rater) behaves like humans is not validated in the experiments. Therefore, we analyzed the correlation between NN-rater's scores for extended reference responses with the human judgments. As this analysis requires human evaluation costs to measure the correlation with each NN-rater for the responses retrieved by each method, we used the BERT-based NN-rater used in υBERTScore, which had the highest correlation with human judgments in the experiments. To obtain human judgments of υBERTScore's extended reference responses, we asked four annotators, who are Japanese native speakers in our research group and were not involved with this study, to evaluate the appropriateness of the extended reference responses given the target utterance. These instructions are identical to the settings in § 5.4; specifically, we used responses of the 15 most similar utterance-response pairs retrieved by υBERTScore for the test data.

Table 6 shows the results of the analysis. First, we observed that the correlations among annotators did not differ from those in Table 4. This implies that the quality of human judgments is almost the same among each other. Second, we confirmed that the BERT-based NN-rater had a high correlation with human judgments. This may be because NN-rater is trained with human-human dialogues and tested in almost the same situation.

Metric	Spearman's ρ			Pearson's r		
	max	min	avg	max	min	avg
BERT-based NN-rater	.533	.508	.584	.498	.490	.556
Human	.775	.670	—	.767	.664	—

Table 6 Correlation between BERT-based NN-rater and human judgment to extended reference responses using BERT; human refers to the inter-rater correlations.

Examples Table 7 shows examples of responses retrieved and evaluated by our method, along with evaluation scores for responses generated by C-BM25. The BLEU score with a single-reference response was almost zero because few words matched in the reference and generated responses. Multi-reference BLEU ($\text{BLEU}_{\text{multi}}$) also scored low. This is because $\text{BLEU}_{\text{multi}}$ uses inappropriate or irrelevant responses as reference responses and is disturbed by them. Conversely, ΔBLEU, υBLEU, and υBERTScore gave a slightly higher score to the generated response, similar to the human evaluation[12]. In ΔBLEU, validation scores are high quality because they are based on human judgment; however, the retrieval method based on exact word matches is poor and often tends to collect inappropriate responses. In Table 7, the low-rated (second) response ("シンクロ率高すぎでしょ": "Synchronization rate is too high") collected only because it matches ("すぎでしょ": "... is too .." or "... is very .."). In contrast, because υBLEU searches for responses based solely on utterance similarity, υBLEU's extended reference responses will have a more diverse word distribution than ΔBLEU's. Finally, υBERTScore also collects various responses and can utilize a BERTScore-based evaluation metric that is calculated by the similarity of the semantic meaning of words.

ΔBLEU and our proposed reference augmentation are useful when the reference and the generated response are not similar in an exact match or even word-based sentence similarity as in this example. In particular, υBERTScore can perform embedding-based flexible evaluation, which captures possible semantically-similar responses, whereas the utterance-based retrieval can collect semantically-dissimilar response candidates.

7 Conclusions

Herein, we propose methods to remove the need for costly human judgment in ΔBLEU (Galley

[12] Note that since evaluation metrics are (meta-)evaluated in correlation with human judgment, it is not possible to decide the metric that is better from this perspective alone.

Utterance:	
puma 描いて一晩経ったらフォロワーが 10 人減っていたので時代はまだ追いついていない (Time has not got me, because my follower reduced by 10 on the next day after I've drawn puma.)	
Reference response:	
おもしろすぎでしょ (It's very funny)	
Extended reference responses by ΔBLEU:	**Averaged human score**
わかりすぎる (I strongly agree with you.)	0.75
シンクロ率高すぎでしょ (Synchronization rate is too high.)	0.125
Extended reference responses by υBLEU:	**NN-rater score**
此れからも素敵な作品楽しみにしてます (I'm looking forward to seeing your nice work.)	0.835
興味は持ったけど dl できないので興味を失いました (I lost interest on it since I couldn't dl it.)	0.523
Extended reference responses by υBERTScore:	**BERT-based NN-rater score**
逆に考えるんだ 濃縮されていると (Think conversely. It's concentrated.)	0.940
むしろここまで来たらとことんまで! (You'd rather go as far as you can!)	0.676
Generated response (score):	
むしろ辞めたほうが良いのでは (You'd better to stop) (human: 0.33, BLEU: 0.01, $\text{BLEU}_{\text{multi}}$: 0.07, ΔBLEU: 0.24, υBLEU: 0.25, υBERTScore: 0.34)	

Table 7 Examples of responses retrieved and evaluated by our method for a given test example, along with evaluation scores for responses generated by C-BM25. BLEU refers to BLEU score with the original response, whereas $\text{BLEU}_{\text{multi}}$ refers to BLEU score with υBLEU's extended references. (Note: $\text{BLEU}_{\text{multi}}$ does not take NN-rater scores into account.) For comparison, we normalized all evaluation scores to the interval for BLEU, i.e., [0, 1].

et al. 2015) and obtain automatic uncertainty-aware metrics for dialogue systems. Our proposed methods rate diverse reference responses retrieved from massive dialogue logs using a neural network trained with automatically-collected training data and score the generated responses using the retrieved responses and their scores. Experimental results on massive Twitter dialogue data revealed that υBLEU is comparable to human-aided ΔBLEU and is effective if it is combined with reference-free metrics in RUBER. We also confirmed that BERT-extended υBLEU, υBERTScore, is the best performance in the experiments. These results confirm the importance of considering the diversity of outputs in evaluating open-domain dialogue systems.

We release all code and datasets (tweet IDs) to promote the reproducibility of our experi-

ments[13]. The readers are referred to our code to evaluate their dialogue systems for their native languages while reusing dialogue data used to train the target dialogue systems for evaluation to run υBLEU.

Limitations

More powerful open-domain dialogue systems such as BlenderBot (Roller et al. 2021) and Meena (Adiwardana et al. 2020) are publicly available, although in this article we evaluated only VHRED, retrieval-based system, and human as a dialogue system. Therefore, the effectiveness of the proposed method for these new powerful dialogue systems has not yet been verified and guaranteed. Because it is difficult for a single research group to perform a comprehensive evaluation of various dialogue systems, our research communities need to adopt automatic metrics along with human judgments to confirm the applicability of the metric. Smith et al. (2022) reported that whether the differences in performance of a dialogue system (BlenderBot) with different parameters are detectable by human judgment depends on the method of human judgment. Therefore, even if an automatic evaluation method is replaced by human judgment, it is necessary to discuss what differences can be detected by the evaluation method. This study merely demonstrates that the proposed method can address the diversity of responses, which is a challenge in open-domain dialogue evaluation research, by augmenting a reference response of the reference-based evaluation metric. Therefore, what differences in evaluation methods are detectable for various dialogue systems should be carefully discussed as a general issue in open-domain dialogue evaluation research.

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR19A4, Japan, and JSPS KAKENHI Grant Number 21H03494 and 22H00508.

References

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. (2020). "Towards a Human-like Open-

[13] http://www.tkl.iis.u-tokyo.ac.jp/%7Etsuta/jnlp

Domain Chatbot." *arXiv preprint arXiv:2001.09977*.

Banerjee, S. and Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches." In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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)*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., and Dolan, B. (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)*, pp. 445–450, Beijing, China. Association for Computational Linguistics.

Ghazarian, S., Wei, J., Galstyan, A., and Peng, N. (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*, pp. 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Gupta, P., Mehri, S., Zhao, T., Pavel, A., Eskenazi, M., and Bigham, J. (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*, pp. 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online. Association for Computational Linguistics.

Hashimoto, T. B., Zhang, H., and Liang, P. (2019). "Unifying Human and Statistical Evaluation for Natural Language Generation." 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), pp. 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Higashinaka, R., Kawamae, N., Sadamitsu, K., Minami, Y., Meguro, T., Dohsaka, K., and Inagaki, H. (2011). "Building a Conversational Model from Two-tweets." In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 330–335.

Ji, T., Graham, Y., Jones, G., Lyu, C., and Liu, Q. (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)*, pp. 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). "Adam: A Method for Stochastic Optimization." In Bengio, Y. and LeCun, Y. (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Kudo, T. and Richardson, J. (2018). "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium. Association for Computational Linguistics.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset." In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Lin, C.-Y. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (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*, pp. 2122–2132, Austin, Texas. Association for Computational Linguistics.

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., and Pineau, J. (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)*, pp. 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Mehri, S. and Eskenazi, M. (2020a). "Unsupervised Evaluation of Interactive Dialog with DialoGPT." In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 225–235, 1st virtual meeting. Association for Computational Linguistics.

Mehri, S. and Eskenazi, M. (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*, pp. 681–707, Online. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." In Bengio, Y. and LeCun, Y. (Eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "Bleu: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. D. (2014). "GloVe: Global Vectors for Word Representation." In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Phy, V., Zhao, Y., and Aizawa, A. (2020). "Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems." In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ritter, A., Cherry, C., and Dolan, B. (2010). "Unsupervised Modeling of Twitter Conversations." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 172–180, Los Angeles, California. Association for Computational Linguistics.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1995). "Okapi at TREC-3." In *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 109–126. Gaithersburg, MD: NIST.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., and Weston, J. (2021). "Recipes for Building an Open-Domain Chatbot." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325, Online. Association for Computational Linguistics.

Sai, A. B., Mohankumar, A. K., Arora, S., and Khapra, M. M. (2020). "Improving Dialog Evalu-

ation with a Multi-reference Adversarial Dataset and Large Scale Pretraining." *Transactions of the Association for Computational Linguistics*, **8**, pp. 810–827.

Sato, S., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M. (2017). "Modeling Situations in Neural Chat Bots." In *Proceedings of ACL 2017, Student Research Workshop*, pp. 120–127, Vancouver, Canada. Association for Computational Linguistics.

Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues." *Proceedings of the AAAI Conference on Artificial Intelligence*, **31** (1).

Shang, L., Lu, Z., and Li, H. (2015). "Neural Responding Machine for Short-Text Conversation." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1577–1586, Beijing, China. Association for Computational Linguistics.

Sinha, K., Parthasarathi, P., Wang, J., Lowe, R., Hamilton, W. L., and Pineau, J. (2020). "Learning an Unreferenced Metric for Online Dialogue Evaluation." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2430–2441, Online. Association for Computational Linguistics.

Smith, E., Hsu, O., Qian, R., Roller, S., Boureau, Y.-L., and Weston, J. (2022). "Human Evaluation of Conversations is an Open Problem: Comparing the Sensitivity of Various Methods for Evaluating Dialogue Agents." In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pp. 77–97, Dublin, Ireland. Association for Computational Linguistics.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 196–205, Denver, Colorado. Association for Computational Linguistics.

Tao, C., Mou, L., Zhao, D., and Yan, R. (2018). "RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems." *Proceedings of the AAAI Conference on Artificial Intelligence*, **32** (1).

Vinyals, O. and Le, Q. V. (2015). "A Neural Conversational Model." *CoRR*, **abs/1506.05869**.

Zhang, C., D'Haro, L. F., Banchs, R. E., Friedrichs, T., and Li, H. (2021). *Deep AM-FM: Toolkit for Automatic Dialogue Evaluation*, pp. 53–69. Springer Singapore, Singapore.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). "BERTScore: Evaluating Text Generation with BERT." In *International Conference on Learning Representations*.

Yuma Tsuta: He received his B.S. in Integrated Science from the University of Tokyo in 2018, and his M.S. in Information and Communication Engineering from the University of Tokyo in 2020. He is currently working towards a doctoral degree at the University of Tokyo. His research interests include natural language processing, and in particular open-domain dialogue systems.

Naoki Yoshinaga: He received his Ph.D. from the University of Tokyo in 2005. He has been JSPS Research Fellow (DC1, PD) from 2002 to 2008, and Associate Professor at the University of Tokyo since 2016. His research interests include computational linguistics and machine learning, in particular efficient and adaptive natural language processing in the wild.

Masashi Toyoda: He received his Ph.D. degree from the Tokyo Institute of Technology in 1999. He worked as Associate Professor at the University of Tokyo from 2006 to 2018, and has been professor since 2018. His research interests include analysis and interactive visualization of Web, social media, and IoT data.

(Received October 1, 2022)
(Revised January 13, 2023)
(Accepted February 20, 2023)