υBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems

Yuma Tsuta
The University of Tokyo
tsuta@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science
The University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Masashi Toyoda
Institute of Industrial Science, The University of Tokyo
toyoda@tkl.iis.u-tokyo.ac.jp

Abstract

Because open-domain dialogues allow diverse responses, basic reference-based metrics such as BLEU do not work well unless we prepare a massive reference set of high-quality responses for input utterances. To reduce this burden, a human-aided, uncertainty-aware metric, ΔBLEU, has been proposed; it embeds human judgment on the quality of reference outputs into the computation of multiple-reference BLEU. In this study, we instead propose a fully automatic, uncertainty-aware evaluation method for open-domain dialogue systems, υBLEU. This method first collects diverse reference responses from massive dialogue data and then annotates their quality judgments by using a neural network trained on automatically collected training data. Experimental results on massive Twitter data confirmed that υBLEU is comparable to ΔBLEU in terms of its correlation with human judgment and that the state of the art automatic evaluation method, RUBER, is improved by integrating υBLEU.

1 Introduction

There has been increasing interest in intelligent dialogue agents such as Apple Siri, Amazon Alexa, and Google Assistant. The key to achieving higher user engagement with those dialogue agents is to support open-domain non-task-oriented dialogues to return a meaningful response for any user input.

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 on 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 made from actual conversations. It is, therefore, hard for these reference-based metrics to consider uncertain responses without writing additional reference responses by hand (§ 2).

To remedy this problem, Galley et al. (2015) proposed ΔBLEU (§ 3), a human-aided evaluation method for text generation tasks with uncertain outputs. The key idea behind ΔBLEU is to consider human judgments on reference responses with diverse quality in BLEU computation. Although ΔBLEU correlates more strongly with human judgment than BLEU does, it still requires human intervention. Therefore it cannot effectively evaluate open-domain dialogue systems in a wide range of domains.

To remove the human intervention in ΔBLEU, we propose an automatic, uncertainty-aware evaluation metric, υBLEU. This metric exploits reference responses that are retrieved from massive dialogue logs and rated by a neural network trained with automatically collected training data (§ 4). We first retrieve diverse response candidates according to the similarity of utterances to which the responses were directed. We then train a neural network that judges the quality of the responses by using training data automatically generated from utterances with multiple responses. We also propose integrating υBLEU into the state of the art evaluation method, RUBER (Tao et al., 2018) (§ 2) to advance the state of the art by replacing its reference-based scorer.

Using our method, we experimentally evaluated responses generated by dialogue systems such as a retrieval-based method (Liu et al., 2016) and a generation-based method (Serban et al., 2017) 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 that correlation (§ 6).

Our contributions are the followings:

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 199–206
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics

- We developed an uncertainty-aware automatic evaluation method for dialogue systems. Our method automates the human ratings required in ΔBLEU while keeping the performance.

- We showed that integrating υBLEU into RUBER greatly improves RUBER's performance by providing the robustness to evaluate responses with uncertainty.

2 Related work

This section introduces recent studies on evaluating open-domain dialogue systems. We focus here on model-agnostic methods than can evaluate the quality of a response for a given utterance.[1]

For evaluation of dialogue systems, researchers have adopted existing evaluation metrics for other text generation tasks such as machine translation and summarization. Unfortunately, reference-based metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) correlate poorly with human judgment on evaluating dialogue systems (Liu et al., 2016). This is because only a few responses, or often only one, can be used as reference responses when actual conversations are used as datasets, even though responses in open-domain dialogues can be diverse (Sato et al., 2017).

To consider uncertain responses in open-domain dialogues, Sordoni et al. (2015) attempted to collect multiple reference responses from dialogue logs for each test utterance-response pair. Galley et al. (2015) improved that 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 Daily Dialogue dataset (Li et al., 2017). Although the last two studies empirically showed that the use of human-rated or -created reference responses in evaluation improves the correlation with human judgment, it is costly to create such evaluation datasets for various domains.

As for evaluation methods, ADEM (Lowe et al., 2017) learns an evaluation model that predicts human scores for given responses by using large-scale human-rated responses that are originally generated by humans or dialogue systems. The drawback of that method is the cost of annotation to train the evaluation model. Moreover, the evaluation model has been reported to overfit the dialogue systems used for generating the training data.

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 by using the cosine similarity of their vector representations, while its unreferenced scorer, trained by negative sampling, evaluates the relevance between an input utterance and a generated response. Ghazarian et al. (2019) showed that use of BERT embedding (Devlin et al., 2019) in pretrained vectors improves the unreferenced scorer but not the referenced scorer in RUBER. the referenced scorer is similar to ΔBLEU in that they both are referenced-based evaluation metrics. We later confirm that the referenced scorer in RUBER underperforms our method, and we thus propose replacing it with our method (§ 5.5).

3 Preliminaries

This section reviews 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).

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 and generated response. Specifically, the score is calculated from a modified n-gram precision p_n and a brevity penalty (BP):

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

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

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

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

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

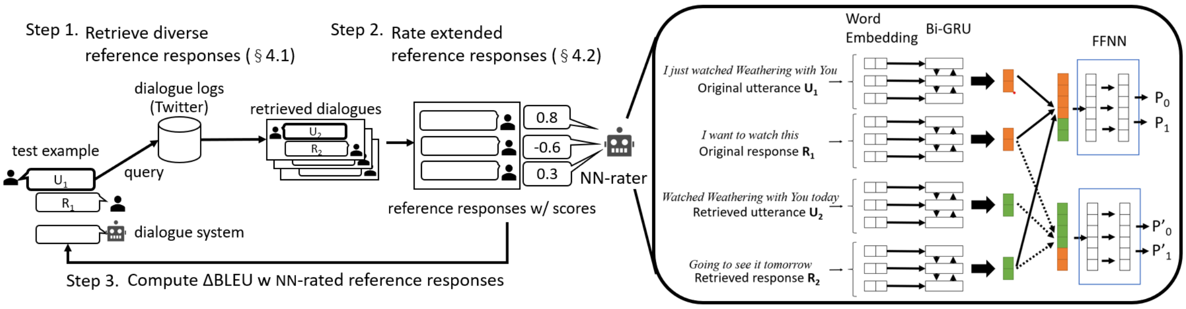

Figure 1: An overview of υBLEU: retrieving diverse reference responses from dialogue logs (§ 4.1) to augment the reference response in each test example, followed by neural network (NN)-rater that judges the their quality (§ 4.2).

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, from Twitter, utterance-response pairs similar to the given pair. The similarities between utterances and between responses are next calculated by using BM25 (Robertson et al., 1994), and they are multiplied to obtain the similarity between the utterance-response pairs. Then, the responses for the top-15 similar utterance-response pairs and the utterance (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:

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

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

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

4 Proposed method: υBLEU

This section describes our approach to the problems 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).

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 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 by using the cosine similarity between utterance vectors obtained from the average of pretrained 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 judges the appropriateness of the extended reference responses for the utterance. To remove this human intervention, we propose rating each reference response by 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. The NN-rater is trained to output the probability that the retrieved response R_2 for

Task (method)	Unit	Training	Validation	Test
response generation	utterance-response pair	2.4M (2018)	10K (2018)	100 (2019)
NN-rater, RUBER	pair of utterance-response pairs	5.6M (2017)	10K (2017)	n/a
reference response retrieval, training for GloVe	utterance-response pair	Approximately 16M (2017)		

Table 1: Statistics of the dialogue data used to run each task. The numbers in the parentheses mean year.

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.

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

We then train the NN-rater in Figure 1 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.

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, $\langle \mathrm{U}_1, \mathrm{R}_1, \mathrm{R}_2 \rangle$ and $\langle \mathrm{U}_2, \mathrm{R}_2, \mathrm{R}_1 \rangle$, as inputs to the NN-rater. Next, we make two vectors by concatenating triplet vectors returned from 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 Bi-GRU to represent a utterance/response vector as $v = [h_f, h_b]$. We then feed each triplet vector to 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 $[-1, 1]$. This formulation is inspired by Tao et al. (2018) and Ghazarian et al. (2019).

5 Experimental Settings

This section describes how to evaluate our method 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 method in terms of its correlation with human judgment (§ 5.3–5.5).

5.1 Twitter dialogue datasets

We built a large-scale Japanese dialogue dataset from Twitter posts of 2.5 million users that have been collected through the user timeline API since March 2011 (Nishi et al., 2016). Posts that are neither retweets nor mentions of other posts were regarded as utterances, and posts mentioning these posts were used as responses.

We use this dataset for training and testing dialogue systems and for training the NN-rater that judges 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 1.

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: 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.

Following Ritter et al. (2010) and Higashinaka et al. (2011), to use a series of dialogues 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 its replies in the obtained dialogues as our dialogue data: 2.4M pairs for training VHRED and for retrieving responses in C-BM25, 10K pairs as validation data for VHRED, and 100 pairs as test data.[2] These dialogues were tokenized with SentencePiece (Kudo and Richardson, 2018) for VHRED and with MeCab 0.996 (ipadic 2.7.0)[3]

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

[3] https://taku910.github.io/mecab/

Metric	Reference retrieval method		Spearman's ρ		Pearson's r	
	Target to compute similarity	Function to compute similarity	max	min	max	min
BLEU	(Only one reference response)		.186	.091	.276	.190
BLEU	Utterance & Response	BM25	.257	.138	.298	.173
BLEU	Utterance only	BM25	.265	.136	.296	.178
BLEU	Utterance & Response	Cosine similarity for GloVe vector	.280	.148	.322	.177
BLEU	Utterance only	Cosine similarity for GloVe vector	**.333**	**.181**	**.366**	**.209**

Table 2: Correlation between human judgment and BLEU with reference responses retrieved by various methods.

for C-BM25 to retrieve responses based on words that are less ambiguous than subwords.

Finally, six Japanese native speakers in our research group evaluated the 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 meaning inappropriate or unrecognizable and 5 meaning very appropriate or seeming to be an actual response.

5.3 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 were set equal in both data. Before these examples were fed to the NN-rater, they are tokenized with SentencePiece.

For the NN-rater, 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 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 1000 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.

5.4 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 (as a parrot return) to apply ΔBLEU and υBLEU. We retrieved utterance-response pairs from approximately 16M utterance-response pairs of our dialogue data (Table 1). 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) from this data.

We then judged the quality of each retrieved reference response by humans for ΔBLEU and by NN-rater for υBLEU in terms of appropriateness as a response to a given utterance. We asked four of the six Japanese native speakers to judge the quality of each retrieved reference response.

5.5 Compared response evaluation methods

We have so far proposed two modifications to improve and automate ΔBLEU: more diverse reference retrieval (§ 4.1) and automatic reference quality judgment (§ 4.2). To see the impact of each modification, we first compare BLEU with various reference retrieval methods. We then compare BLEU with only one reference, ΔBLEU, and υBLEU. Finally, we compared υBLEU with the state of the art evaluation method, RUBER, and examined the performance of RUBER when its referenced scorer was replaced with υBLEU.

Specifically, we applied each evaluation method to the 300 responses (§ 5.2). ΔBLEU and υBLEU used the extended references in evaluation. BLEU used the original (single) references or the extended references. The reference scorer in RUBER used the original (single) references.

Following previous studies (Liu et al., 2016; Tao et al., 2018), 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. To understand the stability of the evaluation, we computed the maximum and minimum correlation with human judgments given by each annotator. All evaluation methods using the modified n-gram precision were calculated with $n \leq 2$ (BLEU-2), following Galley et al. (2015).

6 Results

Table 2 lists the correlations between human judgment and BLEU for each reference retrieval method. In terms of Spearman's ρ, all methods using the extended reference exhibited higher maximum and

Metric	Spearman's ρ		Pearson's r	
	max	min	max	min
ΔBLEU	.366	.300	.360	.294
υBLEU	.330	.281	.394	.332
RUBER				
Unref. & Ref. Scorer	.339	.206	.325	.193
Ref. Scorer only	.188	.071	.075	.016
Unref. Scorer only	.342	.225	.336	.217
Unref. & υBLEU	**.435**	**.323**	**.450**	**.338**
human	.773	.628	.778	.607

Table 3: Correlation between each method and human judgment; human refers to the inter-rater correlations.

minimum correlation with human judgment than BLEU did with only one reference. For Pearson's r, only the proposed retrieval method, which uses an embedding-based similarity for utterances, showed higher minimum correlation than BLEU did with only one reference. This means that the proposed retrieval method was the most appropriate way to extend the reference responses. We, therefore, used reference responses extended by the proposed method for υBLEU in the following evaluation.

Next, Table 3 compares υBLEU with ΔBLEU and the state of the art evaluation method, RUBER. The comparison between υBLEU and BLEU in Table 2 revealed that the use of 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. υBLEU performed better than RUBER did (unreferenced scorer + referenced scorer) for all correlations other than the maximum Spearman's ρ. We attribute the poor performance of RUBER to the poor performance of its referenced scorer, which was even worse than BLEU with only one reference in Table 2. This shows that merely adopting embedding-based similarity does not address the uncertainty of outputs. By replacing the reference scorer in RUBER with our υBLEU, however, we obtained the best overall correlations, which advances the state of the art.

Examples Table 4 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. The υBLEU scores were the closest to human judgment, multi-reference BLEU (BLEU$_{\text{multi}}$) was the secondary closest, and single-reference BLEU was the last.

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:	**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 an interest on it since I couldn't dl it.)	0.523
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, BLEU$_{\text{multi}}$: 0.07, υBLEU: 0.25)	

Table 4: 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, while BLEU$_{\text{multi}}$ refers to BLEU score with the extended references. For comparison, we normalized all evaluation scores to the interval for BLEU, i.e., [0, 1].

7 Conclusions

We have proposed a method to remove the need for costly human judgment in ΔBLEU (Galley et al., 2015) and obtain an automatic uncertainty-aware metric for dialogue systems. Our proposed υBLEU rates diverse reference responses retrieved from massive dialogue logs by using a neural network trained with automatically-collected training data, and it uses the responses and the scores to run ΔBLEU. Experimental results on massive Twitter dialogue data revealed that υBLEU is comparable to human-aided ΔBLEU, and that, by integrating it into RUBER, the state of the art method for evaluating open-domain dialogue systems, we can improve the correlation with human judgment.

We will release all code and datasets (tweet IDs) to promote the reproducibility of our experiments.[4] The readers are referred to our code to evaluate their dialogue systems for their native languages.

Acknowledgments

The research was supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation.

[4] http://www.tkl.iis.u-tokyo.ac.jp/~tsuta/acl-srw-2020/

References

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 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*, pages 103–111, Doha, Qatar. Association for 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.

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.

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.

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.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 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)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Ryuichiro Higashinaka, Noriaki Kawamae, Kugatsu Sadamitsu, Yasuhiro Minami, Toyomi Meguro, Kohji Dohsaka, and Hirohito Inagaki. 2011. Building a conversational model from two-tweets. In *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 330–335.

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

Taku Kudo and John Richardson. 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*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. 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.

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.

Ryosuke Nishi, Taro Takaguchi, Keigo Oka, Takanori Maehara, Masashi Toyoda, Ken-ichi Kawarabayashi, and Naoki Masuda. 2016. Reply trees in Twitter: Data analysis and branching process models. *Social Network Analysis and Mining*, 6(1):26.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 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*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, pages 109–126. National Institute of Standards and Technology (NIST).

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.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Association for the Advancement of Artificial Intelligence*, pages 3295–3301.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 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*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

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.