Designing Precise and Robust Dialogue Response Evaluators

Tianyu Zhao   **Divesh Lala**   **Tatsuya Kawahara**

Graduate School of Informatics

Kyoto University

{zhao,lala,kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

Automatic dialogue response evaluator has been proposed as an alternative to automated metrics and human evaluation. However, existing automatic evaluators achieve only moderate correlation with human judgement and they are not robust. In this work, we propose to build a reference-free evaluator and exploit the power of semi-supervised training and pretrained (masked) language models. Experimental results demonstrate that the proposed evaluator achieves a strong correlation ($>$ 0.6) with human judgement and generalizes robustly to diverse responses and corpora. We open-source the code and data in https://github.com/ZHAOTING/dialog-processing.

1 Introduction

Evaluation of conversational systems has been one major obstacle in dialogue research. Particularly for open-domain dialogues, automated metrics have been shown to correlate poorly with human judgement (Liu et al., 2016). Although human evaluation provides the most accurate assessment, they are slow and expensive. An alternative is to train an evaluator that learns to predict a human-like score. Lowe et al. (2017) proposed ADEM, a supervised regression model, for automatic response evaluation and reported 0.436 Pearson's and 0.428 Spearman's correlations with human judgement. Though better than automated metrics, the scores only indicate moderate correlations. Another criticism from Sai et al. (2019) further pointed out that ADEM produces scores of low deviation and lacks robustness under adversarial attack.

An ideal evaluator should be precise such that its predictions have a strong correlation with human judgement. It should also be robust such

that it generalizes to new dialogues unseen during training. We explored three methods to improve the precision and robustness of response evaluators. 1) We propose building reference-free evaluator since reference-dependent metrics cause the problem of low deviation described by Sai et al. (2019). We also find that the reference-dependent evaluators' performance degrades significantly when we remove ground-truth responses from test data. 2) Tao et al. (2018) proposed an unsupervised model (RUBER) that outperforms supervised ADEM by training on a next sentence prediction (NSP) task. We show that RUBER can be further improved by supervised training on a small amount of annotated data. 3) We make use of strong pretrained models such as RoBERTa (Liu et al., 2019) to obtain better text representations. By combining the three methods, a reference-free, semi-supervised, RoBERTa-based evaluator has better correlation and robustness. Experimental results also show that the model can maintain good performances in cross-domain and low-resource settings.

2 Related Works

Automatic response evaluator was first proposed by Lowe et al. (2017) to mimic human annotator's assessment of response appropriateness. They collected human annotations of response quality for 4,104 context-response pairs, and train a regression network (ADEM) supervisedly by minimizing a squared error. Tao et al. (2018) proposed an unsupervised method (RUBER) to train automatic evaluators, where a model is optimized to distinguish a ground-truth response and a negative-sampling response by minimizing a margin rank loss. This process resembles the next sentence prediction (NSP) task applied in the training of BERT (Devlin et al., 2019). It allows for exploit-

ing a large amount of conversation data and has been shown to outperform ADEM. Using ADEM and RUBER as the baselines of this work, we will analyze their shortcomings and develop solutions to build more precise and robust evaluators.

Next sentence prediction is to predict whether a sentence is a true continuation given a preceding context, where a positive sample is the ground-truth subsequent sentence and a negative sample is a different piece of text. NSP benefits not only evaluation (Tao et al., 2018), but also language understanding (Devlin et al., 2019) and language generation (Bruni and Fernandez, 2017; Wolf et al., 2019).

Dialogue response evaluation can also be improved with better automated metrics and approximation to response quality. Examples of successful attempts to improve automated metrics include exploiting multiple references for comparison (Gupta et al., 2019) and combining human judgement with automated metrics (Hashimoto et al., 2019). Li et al. (2019) demonstrated that single-turn human judgement is not reliable as expected and proposed multi-turn human evaluation. Ghandeharioun et al. (2019) approximated sentiment, semantic similarity, and engagement with new automated metrics and used a hybrid metric in a multi-turn evaluation setting. Dziri et al. (2019) showed that entailment is also an option to approximate dialogue coherence and quality.

3 Background

ADEM is a regression model that takes as inputs a dialogue context vector $\mathbf{c}$, a hypothesis response vector $\hat{\mathbf{r}}$, and a reference response vector $\mathbf{r}$. Its output is the sum of a referenced metric and an unreferenced metric:

$$\text{ADEM}_{ref}(\mathbf{r}, \hat{\mathbf{r}}) = \mathbf{r}^T N \hat{\mathbf{r}}, \quad (1)$$

$$\text{ADEM}_{unref}(\mathbf{c}, \hat{\mathbf{r}}) = \mathbf{c}^T M \hat{\mathbf{r}}, \quad (2)$$

where the encoding vectors are produced by pretrained RNN encoders. M and N are trainable parameters.

RUBER also combines two metrics but computes them differently:

$$\text{RUBER}_{ref}(\mathbf{r}, \hat{\mathbf{r}}) = \frac{\mathbf{r}^T \hat{\mathbf{r}}}{\|\mathbf{r}\| \cdot \|\hat{\mathbf{r}}\|}, \quad (3)$$

$$\text{RUBER}_{unref}(\mathbf{c}, \hat{\mathbf{r}}) = \text{MLP}([\mathbf{c}; \hat{\mathbf{r}}; \mathbf{c}^T M \hat{\mathbf{r}}]; \theta), \quad (4)$$

where $[\cdot; \cdot]$ denotes the concatenation of vectors and MLP is a multi-layer perceptron with nonlinear activation functions. M and θ are trainable parameters.

Besides the differences in metric computation, they are different in training strategy. ADEM uses supervised training to minimize the mean square error between predictions and human scores, while RUBER uses unsupervised training on an NSP task to minimize a margin ranking loss. In Section 5, we combine their advantages to build a better response evaluator.

4 Data Collection

For assessing dialogue response evaluators, we sample 100 dialogues from the test split of the DailyDialog corpus (Li et al., 2017) which contains 13,118 open-domain and human-written conversations. We expand them with extra response hypotheses and collect human annotations of response quality.

Collection of Extra Responses. Besides the ground-truth response, we add responses from different sources for each dialogue context, including 1) a negative-sampling response randomly selected from a different dialogue and 2) responses generated by generative models trained on the training split. We combine 6 generative models (S2S (Sutskever et al., 2014), attentional S2S, HRED (Serban et al., 2016), VHRED (Serban et al., 2017), GPT2-*sm*, and GPT2-*md* (Wolf et al., 2019)) with 3 decoding methods (greedy decoding, ancestral sampling, and nucleus sampling (Holtzman et al., 2019)). The resulting response pool for each dialogue context contains 20 responses of various qualities.

Collection of Human Annotations. From the 2,000 dialogue-response pairs, we select 900 of them and ask Amazon Mechanical Turk workers to rate response *appropriateness* on a 5-point Likert scale. Each pair is rated by four workers. After removing annotation outliers for each pair (Leys et al., 2013), the remaining data reaches good reliability regarding an inter-annotator agreement with Krippendorff's $\alpha > 0.8$ (Krippendorff, 2018).[1] We make a 0.8:0.1:0.1 split of the annotated data for training, validation and test.

Figure 1(a) shows the overall distribution of 900 human scores on response *appropriateness*, and

---

[1]More details of inter-annotator agreement and outlier removal are provided in Appendix A.

(a) Overall score distribution.

(b) Box plot of scores for each response source. GT - ground-truth, NS - negative-sampling.

Figure 1: Distributions of human annotations on response *appropriateness* (§4).

Model	Full Test Data (90 responses)			Excluding Ground-truth (77 responses)		
	Pearson	Spearson	SD	Pearson	Spearson	SD
ADEM						
full	0.34**	0.36**	0.51	0.25	0.23	0.30
ref.	0.32*	0.35**	0.52	0.21	0.23	0.30
unref.	0.26	0.26	0.32	0.28	0.27	0.33
RUBER						
full	0.37**	0.31*	0.67	0.43**	0.41**	0.68
ref.	0.32*	0.29*	0.07	0.12	0.13	0.04
unref.	0.35**	0.29*	1.32	0.43**	0.39**	1.35
Human	1.0	1.0	1.42	1.0	1.0	1.40

Table 1: Comparison between referenced metric and unreferenced metric on the full test data and the ground-truth response-excluded test data (§5.1). **SD** is short for standard deviation. * denotes scores that have p-values < 0.01. ** denotes scores that have p-values < 0.001.

Figure 1(b) shows box plots of human scores for different response sources. The distributions suggest that the created data consists of diverse responses.

5 Methodology

5.1 Reference-free Evaluation

Sai et al. (2019) proved theoretically that the comparison with reference response in the referenced metric causes ADEM to make conservative predictions where scores have a very low standard deviation. To investigate the effect of removing reference from computation, we experiment with the full ADEM and RUBER as well as their referenced and unreferenced versions. As shown in Table 1, the referenced metrics of ADEM and RUBER have much lower standard deviations than human scores. ADEM's unreferenced metric has low scores in both correlation and standard deviation because the full ADEM model is heavily affected by its referenced metric while its unreferenced metric is not fully utilized, especially in the

data set that includes ground-truth responses.

Another important finding is that the referenced metrics' correlations degrade significantly when we remove ground-truth responses from the test data. It suggests that referenced metrics may help evaluators to distinguish a ground-truth response from a non-ground-truth response easily, but they cannot distinguish a good response from a bad one among non-ground-truth responses.

Based on the results, we propose to build reference-free evaluators and avoid direct comparison with reference responses to improve its robustness and diversity.

5.2 Semi-supervised Training

ADEM is a supervised model that relies on human annotations. However, it is expensive to collect large-scale annotated data; On the other hand, RUBER has been shown to reach reasonable correlation scores via only unsupervised training on an NSP task. A natural idea is to apply unsupervised training first and then finetune an evaluator using a

28

Model	Pr.	Spr.	Training data
RUBER			
sup.	0.37**	0.31*	130k
semi-sup.	0.45**	0.41**	130k+720

Table 2: Comparison between original unsupervised RUBER and semi-supervised RUBER (§5.2). **Pr.** and **Spr.** are short for Pearson's correlation and Spearman's correlation, respectively.

relatively small amount of annotated data. Taking RUBER as an example, by finetuning RUBER on 720 annotated samples, we improve its Pearson's correlation from 0.37 to 0.45 and Spearman's correlation from 0.31 to 0.41.

5.3 Powerful Text Encoder

All the metrics mentioned before are based on encoding vectors $\mathbf{r}$, $\hat{\mathbf{r}}$ and $\mathbf{c}$, so a powerful text encoder is essential to building a good evaluator. ADEM and RUBER are both initialized with pretrained RNN response generators. As an alternative, pretrained (masked) language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) can be used as a powerful text encoder and have benefited most downstream tasks in natural language processing (Huang et al., 2019; Lan et al., 2020; Joshi et al., 2020; Shimanaka et al., 2019). We choose RoBERTa-*large* to build our response evaluator.

A RoBERTa evaluator produces an encoding vector $\mathbf{d}$ given a context c and a response $\hat{r}$ and then finally calculates its score via an MLP with a sigmoid function. We rescale the score to match annotator's scale of [1, 5]:

$$\mathbf{d} = \text{RoBERTa}([c; \hat{r}]; \phi), \qquad (5)$$
$$\text{RoBERTa-eval}(c, \hat{r}) = 4 \cdot \text{MLP}(\mathbf{d}; \theta) + 1, \quad (6)$$

where RoBERTa's parameter ϕ and MLP's parameter θ can both be optimized during training.

6 Experimental Evaluations

Table 3 shows the correlation scores and standard deviations of four metric groups. The first group is automated metrics that are based on n-gram overlapping (BLEU-2) or word embedding similarities (Average, Extrema, and Greedy). The second group is the baseline ADEM and RUBER. The third group is the semi-supervised full RUBER model, the semi-supervised unreferenced RUBER model, and the RoBERTa-based evaluator

Model	Pr.	Spr.	SD
Automated Metrics			
BLEU-2	0.31	0.23	0.31
Average	0.25	0.23	0.19
Extrema	0.26	0.26	0.23
Greedy	0.25	0.23	0.21
Baseline Evaluator			
ADEM	0.34**	0.36**	0.51
RUBER	0.37**	0.31*	0.67
Proposed Evaluator			
RUBER			
semi-sup.	0.45**	0.41**	0.42
unref.+semi-sup.	0.43**	0.39**	0.83
RoBERTa-eval	0.64**	0.66**	1.26
Human Judgement			
Human	1.0	1.0	1.42

Table 3: Performances of automated metrics, baseline evaluators, and proposed evaluators (§6).

that combines the three proposed methods. Human scores are given in the final group. Semi-supervised training yields improvement in correlations, and abandoning referenced metrics makes predictions less conservative. The RoBERTa evaluator outperforms the baselines by a large margin and has a much human-like score diversity.

6.1 Transferability Study

We are interested in applying a trained response evaluator to new data of different domains or styles. Therefore, we carry out experiments to study the transferability of the RoBERTa evaluator. In addition to the DailyDialog (DD) corpus, we further collect annotations on 900 responses from the PersonaChat (PC) corpus (Zhang et al., 2018) following the same procedure in Section 4. The evaluator turns out to generalize to a new corpus much better than the baseline RUBER according to results in Table 4. The evaluator trained on the DD corpus achieves even higher correlation scores when applied to the PC corpus. However, performance degradation is observed when applying the evaluator trained on the PC corpus to the DD corpus. It suggests that we should make a careful choice of training data when planning to evaluate our models on different corpora.

6.2 Low Resource Study

Although only 720 annotated samples are used in the experiments above, we explored the possibility

Corpus		Correlation	
Train	Test	Pr.	Spr.
RoBERTa evaluator			
DD	DD	0.64**	0.66**
DD	PC	0.69**	0.69**
PC	PC	0.75**	0.76**
PC	DD	0.50**	0.47**
RUBER			
DD	DD	0.37**	0.31*
DD	PC	0.12	0.17
PC	PC	0.58**	0.57**
PC	DD	0.06	0.06

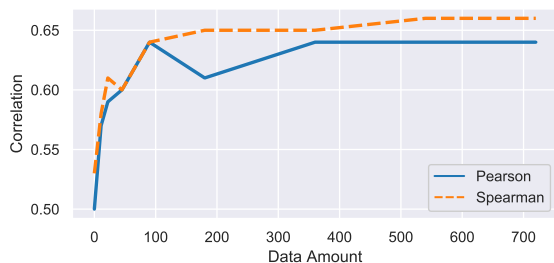Table 4: Correlations of RoBERTa evaluator and RU-BER using training and test data from different corpora (§6.1).

Figure 2: Performance of the RoBERTa evaluator w.r.t amount of supervised training data (§6.2).

of training with even fewer data. Figure 2 shows that, with only around 100 samples, the RoBERTa evaluator can reach performance close to the result obtained using the entire 720 samples.

6.3 Robustness Evaluation

In this section, we address Sai et al. (2019)'s requirements towards a robust evaluator.

1. Not be heavily influenced by the reference response. The proposed evaluator is entirely independent of references.

2. Generalizing to diverse responses. 1) After removing ground-truth from the test data, the RoBERTa evaluator still achieves 0.62 Pearson's correlation and 0.64 Spearman's correlation. 2) The evaluator achieves good performances on diverse responses (see §4) and different corpora (see §6.1).

3. Sensitivity to grammar and relevance of the response. We also collected annotations for *relevance* and *grammatical correctness*. The RoBERTa evaluator trained on *appropriateness* annotations can achieve 0.68 Pearson's and 0.67

Spearman's correlations with *relevance* annotations, while its correlation scores with *grammatical correctness* are only 0.09 and 0.15. However it is understandable because responses of perfect grammar can still be inappropriate in a certain context and grammar itself is not highly correlated with *appropriateness*.[2]

4. Robust against fooling attacks. Unlike in Sai et al. (2019), we have not found any magic responses that can fool the evaluators to output high scores constantly.

7 Conclusion

Automatic dialogue response evaluators have problems in robustness and correlation with human judgement. We investigated three methods to alleviate them: 1) using reference-free metrics, 2) applying semi-supervised training, and 3) exploiting powerful pretrained text encoders. Experimental results demonstrated that our proposed evaluator achieved strong correlation (> 0.6) with human judgement and showed robustness in dealing with diverse responses and a new domain. It can also be trained efficiently with less than 100 annotated samples.

Acknowledgments

References

Elia Bruni and Raquel Fernandez. 2017. Adversarial evaluation for open-domain dialogue generation. In *SIGDIAL 2017, The 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 284–288.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019, The 2019 Conference of the North American Chapter of the Association for*

---

[2]According to the collected annotations, *appropriateness* and *relevance* are highly correlated with 0.91 Pearson's and 0.91 Spearman's scores, while *appropriateness* and *grammatical correctness* have only 0.37 Pearson's and 0.34 Spearman's scores.

Computational Linguistics: Human Language Technologies, pages 4171–4186.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *NAACL HLT 2019, The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3806–3812.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *NeurIPS 2019, Advances in Neural Information Processing Systems 32*, pages 13658–13669.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *SIGDIAL 2019 Workshop, The 20th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 379–391.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *NAACL HLT 2019, The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1689–1701.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *ICLR 2020, The 5th International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *EMNLP-IJCNLP 2019, The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2391–2401.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *ICLR 2020, The 5th International Conference on Learning Representations*.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP 2017, The 8th International Joint Conference on Natural Language Processing*, volume 1, pages 986–995.

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 *EMNLP 2016, The 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 *ACL 2017, The 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1116–1126.

Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. In *AAAI 2019, The 33rd AAAI Conference on Artificial Intelligence*, pages 6220–6227.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *The 30th AAAI Conference on Artificial Intelligence*, volume 16, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI 2017, The 31st AAAI Conference on Artificial Intelligence*, pages 3295–3301.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine translation evaluation with bert regressor. *arXiv preprint arXiv:1907.12679*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS 2014, Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI 2018, The 32nd AAAI Conference on Artificial Intelligence*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

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

Tianyu Zhao and Tatsuya Kawahara. 2019. Effective incorporation of speaker information in utterance encoding in dialog. *arXiv preprint arXiv:1907.05599*.

A Inter-annotator Agreement and Outlier Removal

In the process of collecting human annotations (§4), we collect 3,600 scores in total from 185 Amazon MTurk workers (4 scores for each context-response pair). To assess the data's reliability, we use the Krippendorff's α (Krippendorff, 2018) instead of commonly used Cohen's κ and Fleiss' κ, because Krippendorff's α can handle 1) an arbitrary number of annotators, 2) various levels of measurement (e.g. nominal, interval), and 3) missing data.

The Krippendorff's α of the original 3,600 annotations of response *appropriateness* is 0.431, which is considered not good according to the interpretation of the number in Table 5. Therefore, we decided to remove the outliers to improve the inter-annotator agreement. We detected outliers for each of the 900 four-annotation groups using the median absolute deviation (MAD) method (Leys et al., 2013). By setting the deviation threshold as 1.0, we identified 895 annotations as outliers. On the remaining 2,705 annotations (roughly 1 annotation is removed for each group), the Krippendorff's α reaches 0.815, which suggests that the data is reliable for the subsequent experiments.

B Experimental Settings

The ADEM and RUBER models use a 2-layer bidirectional gated recurrent unit (BiGRU) sentence encoder with 500 hidden units and a 2-layer BiGRU dialogue encoder with 500 hidden units. The encoders are initialized with the parameters of a pretrained HRED's encoders of the same architecture. To encode speaker information, we concatenate each sentence embedding with a 30-dimensional speaker embedding that indicates whether the sentence's speaker is identical to the response's speaker (Zhao and Kawahara, 2019). Principal component analysis (PCA) is applied to project response and context embeddings into low-dimensional vectors in ADEM. The number of principal components is 50. The RoBERTa evaluator is based on a pretrained RoBERTa-*large* model, and we finetune the entire model in our experiments.

Table 6 shows the hyper-parameters in unsupervised training and supervised training. Following the original paper, we freeze the ADEM's encoders and only finetune its parameters M and N, and thus a larger learning rate is used for ADEM. In all experiments, we decay the learning rate with a 0.1 decay rate when a model's validation loss does not improve and stop training early if the learning rate is less than 1e-7.

C Model Output Distributions

The distribution of human annotation scores on the 900 annotated responses has been given in Figure 1(a). To analyze the distribution of model outputs, we show the distributions of human annotation, ADEM's outputs, RUBER's outputs, and RoBERTa-eval's outputs on the test data of 90 responses in Figure 3. We found that: 1) The distribution of human score is similar to that in Figure 1(a). 2) The proposed RoBERTa evaluator's output has a flatter distribution than human scores. 3) The baseline RUBER and ADEM both have very peaky pseudo-Gaussian distributions whose means are around 3.

D Robustness to Changes in Input and Output

We conduct two sets of experiments to see whether the RoBERTa evaluator's performance would be affected by a slight change in its input and output.

Adding Gaussian Noise to Input. We added Gaussian noise ($\mu = 0.0$) to human annotations

Krippendorff's α	Interpretation
<0.67	not good
0.67~0.8	allowing tentative conclusions to be drawn
>0.8	good reliability

Table 5: Interpretation of Krippendorff's α. (§A)

Hyper-parameter	ADEM	RUBER	RoBERTa-eval
Unsupervised Training			
learning rate		1e-4	3e-6
batch size		30	3
epochs		30	2
Supervised Training			
learning rate	1e-3	1e-4	3e-6
batch size	30	30	3
epochs	50	50	50

Table 6: Optimization hyper-parameters.

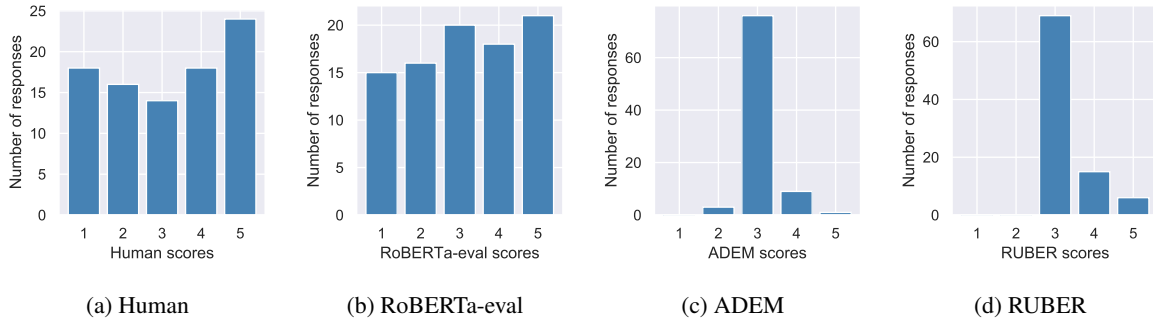(a) Human   (b) RoBERTa-eval   (c) ADEM   (d) RUBER

Figure 3: Distributions of human annotations and model outputs on the test data (90 responses).

and ran 100 trials with random seeds from 1 to 100. With $\sigma = 0.1$, the RoBERTa evaluator's performance doesn't change much (Pearson's correlation from 0.64 to 0.64, Spearman's correlation from 0.66 to 0.65). With $\sigma = 0.5$, the performance degrades more (Pearson's correlation from 0.64 to 0.61, Spearman's correlation from 0.66 to 0.62). Considering that 0.5 σ is high and may skew the original human judgement, we believe the evaluator is not greatly affected by the noise.

Discretizing Output. We also tried discretizing the evaluator's outputs (from [1, 5] to {1, 2, 3, 4, 5}) and observed a minimal improvement (Pearson's correlation from 0.64 to 0.65, Spearman's correlation from 0.66 to 0.66). Generally speaking, there is no dramatic change in the model's performance when we apply these transformations to the output scores. We believe this shows our model to be fairly robust.