MME-CRS: Multi-Metric Evaluation Based on Correlation Re-Scaling for Evaluating Open-Domain Dialogue

Pengfei Zhang[1], Xiaohui Hu[1], Kaidong Yu[1], Jian Wang[1], Song Han[2], Cao Liu[1*], Chunyang Yuan[1]

[1]Meituan

[2]Beijing University of Posts and Telecommunications

{zhangpengfei36, huxiaohui08, yukaidong, wangjian79, liucao, yuanchunyang}@meituan.com, hsong@bupt.edu.cn

Abstract

Automatic open-domain dialogue evaluation is a crucial component of dialogue systems. Recently, learning-based evaluation metrics have achieved state-of-the-art performance in open-domain dialogue evaluation. However, these metrics, which only focus on a few qualities, are hard to evaluate dialogue comprehensively. Furthermore, these metrics lack an effective score composition approach for diverse evaluation qualities. To address the above problems, we propose a Multi-Metric Evaluation based on Correlation Re-Scaling (MME-CRS) for evaluating open-domain dialogue. Firstly, we build an evaluation metric composed of 5 groups of parallel sub-metrics called Multi-Metric Evaluation (MME) to evaluate the quality of dialogue comprehensively. Furthermore, we propose a novel score composition method called Correlation Re-Scaling (CRS) to model the relationship between sub-metrics and diverse qualities. Our approach MME-CRS **ranks first** on the final test data of DSTC10 track5 subtask1 "Automatic Open-domain Dialogue Evaluation Challenge" with a large margin, which proved the effectiveness of our proposed approach.

Introduction

Automatic evaluation is a crucial component for the development of open-domain dialogue systems (Danescu-Niculescu-Mizil and Lee 2011; Yao et al. 2017). The goal of dialogue evaluation is to produce evaluation scores that correlate well to human judgments (scores) over multiple dialogue qualities, i.e., fluency, relevancy, and specificity (Zhang et al. 2018; Weston, Dinan, and Miller 2018). In Dialogue System Technology Challenge 10 (DSTC10) [1], the track5 sub-task1 "Automatic Open-domain Dialogue Evaluation" (Chen et al. 2021) proposes such a challenge in which all participants need to seek effective automatic dialogue evaluation metrics on 14 datasets (37 evaluation qualities in total) during the development phase. A single overall score must be submitted for each dialogue in test datasets during the final evaluation phase.

Nowadays, word overlap-based metrics and embedding-based metrics are standard automatic evaluation metrics. Word overlap-based metrics, which measure the overlapping words between reference and candidate responses, have been used to evaluate the dialogue responses (Papineni et al. 2002; Banerjee and Lavie 2005; Sordoni et al. 2015). Embedding-based metrics measure the evaluation quality of a response by calculating the semantic similarity between the model response and corresponding reference, such as Greedy Matching (Rus and Lintean 2012), Embedding Averaging (Wieting et al. 2016), and BERTScore (Zhang et al. 2019).

However, the aforementioned metrics heavily rely on the given references, and it has been shown that these metrics are ineffective due to the one-to-many nature of dialogue (Zhao, Zhao, and Eskenazi 2017). Recently, learning-based metrics, which aim to predict the scores of various qualities of response, have a better correlation with human judgment (Tao et al. 2018; Ghazarian et al. 2019; Lan et al. 2020). For example, USL-H (Phy, Zhao, and Aizawa 2020) designs BERT-based (Devlin et al. 2019) classifiers for three groups of evaluation qualities: understandability (Nübel 1997), sensibleness (Adiwardana et al. 2020), and likability. USL-H also applies a simple weighted sum to integrate the scores of each evaluation quality. Therefore, USL-H achieves good correlations with human judgment.

Nevertheless, these metrics have some important issues in dealing with dialogue evaluation tasks. First, most evaluation models, which only focus on a few evaluation aspects, are difficult to fully measure the quality of the open-domain dialogue. For example, USL-H ignores some important qualities like topic transition dynamics (Huang et al. 2020) and user engagement in dialogue (Ghazarian et al. 2020). Second, these metrics lack an effective score composition approach to integrate scores generated for each evaluation quality.

To address the above issues, we propose a Multi-Metric Evaluation based on Correlation Re-Scaling (MME-CRS) for evaluating open-domain dialogue as follows. Firstly, to evaluate the dialogue quality more comprehensively, we design 5 groups of sub-metrics for sub-task1 instead of three groups of metrics designed by USL-H. Second, we propose a novel score composition method called Correlation Re-Scaling (CRS) to composite metric scores. Our proposed approach ranks first and achieves an average Spearman correlation score of 31.04% on the test dataset, which is 1.11% higher than the second.

In particular, we summarize our contributions in this pa-

[1]https://dstc10.dstc.community/home

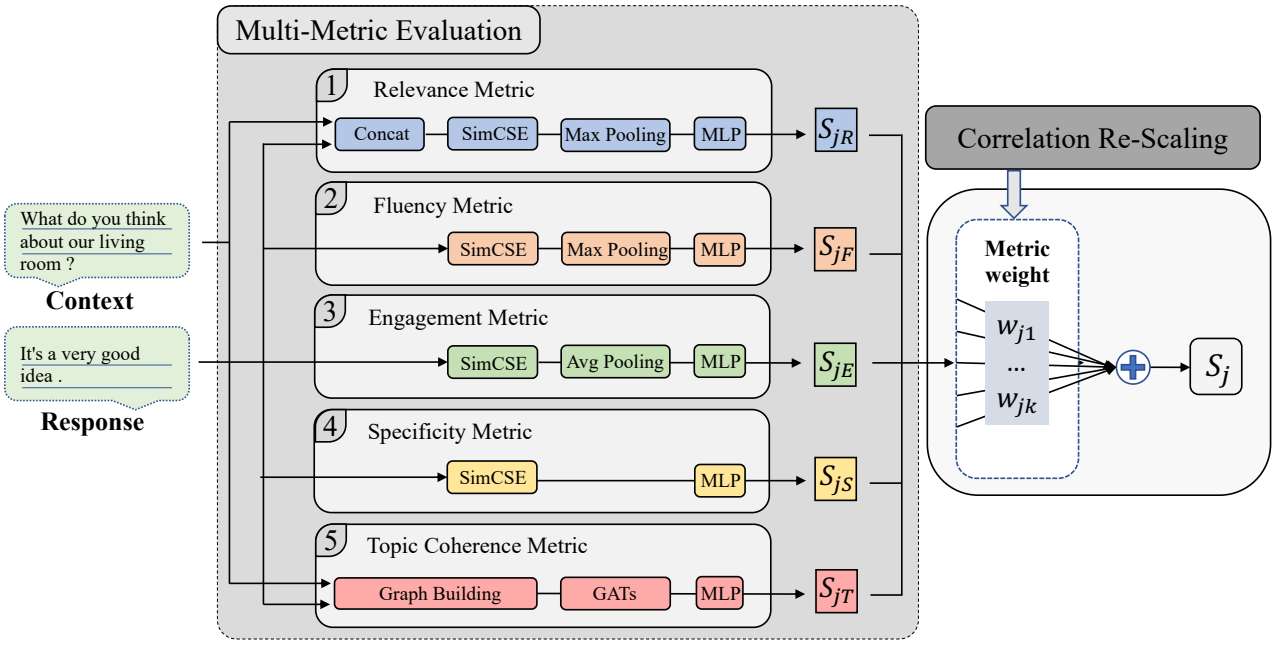

Figure 1: The architecture of the proposed evaluation approach. The "metric weight" is pre-computed using our CRS method. The S_j is the final composition score of evaluation quality q_j for the context-response pair in Figure 1.

per as follows:

- We design an evaluation metric composed of 5 groups of sub-metrics to better evaluate the comprehensive quality of open-domain dialogue.
- We propose a novel score composition method CRS to integrate sub-metric scores more effectively. The weight distribution generated by CRS generalizes well on unseen test data.
- Our proposed metric MME-CRS ranks first on the "Automatic Open-domain Dialogue Evaluation" of DSTC10 track5 task1 with a large margin, which proves the superiority of our designed metrics and CRS method.

Methodology

Figure 1 shows the architecture of our proposed metric MME-CRS. In this section, we will first introduce 5 groups of sub-metrics in detail. Then score composition approach CRS is discussed to integrate sub-metric scores for diverse qualities.

Automatic Evaluation Metrics

The evaluation quality contains various aspects, such as fluency, relevancy, specificity, and user engagement. For example, sub-task1 of DSTC10 track5 contains 14 development datasets, of which 37 different qualities are included in the total. What's more, the evaluation of each aspect usually relies on several metrics, and the weight distribution over sub-metric varies from aspect to aspect. To better measure each evaluation aspect of dialogue, we design 5 groups of fundamental sub-metrics as follows.

Fluency Metric (FM) quantifies whether or not a response is fluency or understandable. A fluency utterance does not have to be grammatically correct because an open-domain response is usually the central part of a complete sentence. The auxiliary verb or stop words may be missing.

We use this characteristic to build a training set of fluent and non-fluent responses. First, we randomly determine if a response r is fluent. If it is, we assign response r with label one and randomly apply one of the following rules: (i) no modification, (ii) delete each stopword with a probability of 0.5. Otherwise, we label response r with zero and apply one of the following rules following Sinha et al. (2020) for negative sampling: (i) word reorder (shuffle the order of all words), (ii) word drop (randomly drop x% words), or (iii) words repeat (randomly select span(s) of words and randomly repeat them).

For a response r with $(w_1, w_2, ...w_n)$ words, we fine-tune SimCSE (Gao, Yao, and Chen 2021) to embed each word in r and apply Max-Pooling to get the utterance embedding. Then a Softmax layer is used to obtain the probability, and we use it as the fluency score S_F.

Relevance Metric (RM) measures coarse-grained relevance between context and response. We fine-tune another SimCSE model based on the next utterance prediction task to predict whether a context-response pair is relevant or not. Similar to the fluency metric, we first randomly determine a context-response pair from the Daily Dialog dataset (Li et al. 2017) is valid or not. For the valid case, we randomly apply one of the following changes to the response: (i) no modification, (ii) remove stop words.

Lan et al. (2020) observes that most random sampled negative responses are low-quality, and the decision boundary

learned is far from the actual decision boundary, which hurts the performance. Hence, for the invalid case, we propose a simple but effective negative response sampling method. First, we randomly choose ten responses from the response pool and compute the Word2Vec similarity (Mikolov et al. 2013) between reference and candidate responses. Then we sort candidate responses based on their similarity score and choose the middle one as a negative response.

To fine-tune the SimCSE model, we first concatenate a context-response pair to a single sentence. Then we compute the score S_R using the same approach as the fluency metric.

Topic Coherence Metric (TCM) measures fine-grained topic transition dynamics of dialogue flows. Huang et al. (2020) demonstrates the effectiveness of incorporation graph information into dialogue evaluation. Following Huang et al. (2020), topic-level dialogue graphs are firstly constructed based on ConceptNet (Speer, Chin, and Havasi 2017). The topic transition dynamics over topic-level dialogue graphs are modeled applying a graph neural network. Then the topic-level graph representation is fed into an MLP layer to predict topic coherence score S_T. Huang et al. (2020) also embeds the context-response pair and jointly predict coherence score together with topic-level graph representation. The former embedding is ignored in this part to focus on the topic coherence metric.

Engagement Metric (EM) measures whether the user is willing to participate in the dialogue. We build a training set based on the human engagement scores. User engagement score usually ranges from 0 to 5, and the user's enthusiasm is proportional to the engagement score. Ghazarian et al. (2020) propose to label response with engagement score less than two as zero, while we find that scaling the engagement score to between 0 and 1 yields more significant benefits.

We train an utterance-level engagement classifier to predict whether the user engagement is high or low. Specifically, for a response r with $(w_1, w_2, ..., w_n)$ words, we fine-tune SimCSE to get the contextual embedding h_i for each word w_i. We use average-pooling here to get the embedding of the whole response. Then an MLP layer followed by a Softmax layer is added to predict the engagement score S_E.

Ghazarian et al. (2020) aggregates the embedding of both context and response to predict the score of user engagement, while we observe that user engagement mainly relies on the model response. The relationship between dialogue context and response should be handled by relevancy metric or topic coherence metric.

Specificity Metric (SM) measures the model's ability to handle diverse words in complex open-domain talking context. We introduce the specificity metric here because some deep models tend to generate general or ambiguous answers.

Mehri and Eskenazi (2020b) uses a Roberta model (Liu et al. 2019) to compute the mask language model (MLM) task, while we use a more light SimCSE model following other proposed sub-metrics. Similar to (Phy, Zhao, and Aizawa 2020), we only use the response r with $(w_1, w_2, ..., w_n)$ words to compute specific score. In detail, we mask each response word w_i and predict negative log-likelihood (SM-LL) based on SimCSE-MLM. We also investigate negative cross-entropy (SM-NCE) and perplexity (SM-PPL) to further improve the effectiveness of specific metrics.

Correlation Re-Scaling Method

Instead of designing a score composition function for the overall aspect alone, we propose to compute weight distribution along designed sub-metrics for each evaluation aspect. The evaluation of each evaluation aspect usually relies on several designed sub-metric. For example, suppose an annotator thinks a response generated by the dialogue model is specific. In that case, he probably implies that the response is also fluent and relevant to the dialogue context. However, the designed specific metric is only trained to predict the specific score for a response. Hence, to better evaluate each dialogue aspect, we propose to model the relationship between designed sub-metrics and diverse evaluation qualities.

We propose a novel Correlation Re-Scaling (CRS) method to compute the weight distribution for each aspect. For a dialogue evaluation dataset D_i with $(q_1, q_2, ..., q_n)$ qualities, we first randomly sample 300 dialogues for Spearman correlation computation.

For each dialogue q_{ij} in dataset D_i, we compute fundamental sub-metric scores as S_{ijk}, where k is the number of sub-metrics. If Spearman correlation score S_{ijk} is less than 0, then the corresponding sub-metric is believed to have no contribution to dialogue quality q_{ij}; thus, S_{ijk} is simply set to 0. We treat correlation score S_{ijk} as the importance of the corresponding sub-metric to quality q_{ij}.

We believe that important sub-metrics should be given higher weight, it is significant for score composition over multiple scores. Hence we compute the normalized weight distribution w_{ijk} as follows:

$$w_{ijk} = \frac{S_{ijk}^{d_{ij}}}{\sum_k S_{ijk}^{d_{ij}}} \tag{1}$$

Where d_{ij} is the power number of S_{ijk}, and the assigned weight to S_{ijk} is $S_{ijk}^{d_{ij}-1}$. The larger d_{ij} is, the more weight is given to more important sub-metrics. According to our experiments, the effect of the score composition method works best when $max(S_{ijk})$ is between 1/3 and 1/2. It is a simple but effective way to determine the value of d_{ij}.

To further improve the generalization ability of CRS method, we calculate the average w_{ijk} of 14 development datasets as follows:

$$w_{jk} = \frac{1}{|D_{q_j}|} \sum_i w_{ijk} \tag{2}$$

Where $|D_{q_j}|$ is the number of development datasets that have q_j quality, and w_{jk} is the normalized weight distribution over each sub-metric for diverse qualities.

For each test dataset D_i, we first compute 7 kinds of sub-metric scores. Then the composition score for each evaluation quality q_{ij} can be computed as follows:

$$S_{ij} = \sum_k w_{jk} \cdot S_{ijk} \tag{3}$$

Dataset	Turns	Qualities	Annos
DSTC6	40000	1	400000
DSTC7	9900	1	29700
Persona-Chatlog	3316	9	29844
PersonaChat-USR	300	6	5400
TopicalChat-USR	360	6	6480
FED-Turn	375	9	3348
FED-Conversation	125	11	1364
DailyDialog-Gupta	500	1	1500
DailyDialog-Zhao	900	4	14400
PersonaChat-Zhao	900	1	3600
DailyDialog-Grade	300	1	3000
Empathetic-Grade	300	1	3000
ConvAI2-Grade	300	1	3000
HUMOD	9500	2	57000

Table 1: Development dataset statistics of the DSTC10 track5 task1. The Turns column is the number of dialogue in the dataset, and the Annos column is the total number of human annotations.

Dataset	Turns	Qualities	Annos
JSALT	741	1	2822
ESL	1242	1	1242
NCM	2461	1	2461
DSTC10-Topical	4500	4	72000
DSTC10-Persona	5000	4	91360

Table 2: Test dataset statistics of the DSTC10 track5 task1.

Experiments

Datasets

Dataset for Metric Training. The organizers of the task1 require that the development datasets are only allowed for validating the proposed metrics, not for training the evaluation systems.

Hence, we select the Daily Dialog dataset (Li et al. 2017) for training our metrics (except for the user engagement metric). The Daily Dialog dataset, which is about day-to-day communication on daily topics, consists of 11118/1000/1000 dialogues for train/valid/test sets, respectively. As for the user engagement metric, we use 13124 responses from ConvAI datasets[2] and scale the engagement score proportionally to between 0 and 1.

Development dataset. During the development phase, the evaluation metrics need to evaluate on 14 datasets (37 evaluation qualities in total). The organizers of "Automatic Open-domain Dialogue Evaluation" identify the following datasets to test the effectiveness of the proposed evaluation metrics [3]:

- DSTC6 (Hori and Hori 2017).
- DSTC7 (Galley et al. 2019).
- Persona-Chatlog (See et al. 2019).
- PersonaChat-USR (Mehri and Eskenazi 2020b).
- TopicalChat-USR (Mehri and Eskenazi 2020b).
- FED-Turn (Mehri and Eskenazi 2020a).
- FED-Conversation (Mehri and Eskenazi 2020a).
- DailyDialog (Gupta et al. 2019).
- DailyDialog (Zhao, Lala, and Kawahara 2020).
- PersonaChat (Zhao, Lala, and Kawahara 2020).
- DailyDialog (Huang et al. 2020).
- Empathetic (Huang et al. 2020).
- ConvAI2 (Huang et al. 2020).
- HUMOD (Merdivan et al. 2020).

Test dataset. During the final test phase, 5 datasets (including 11 evaluation qualities in total) are introduced by task1 organizers to fully evaluate the proposed metrics. The dataset statistics are listed in Table 2.

Evaluation Criteria

The Spearman correlation is used in the "Automatic Open-domain Dialogue Evaluation Challenge" of DSTC10. The Spearman correlations between the submitted scores and corresponding human scores will be computed per evaluation category per dataset. The submissions from different participants will be ranked by the average correlation scores across all the datasets' evaluation qualities.

Training Details

We use a pre-trained SimCSE model (Gao, Yao, and Chen 2021) to fine-tune proposed metrics except for the topic coherence metric, and the model weights of different metrics are not shared. All the metrics based on the SimCSE model are trained with an Adam optimizer (Kingma and Ba 2015) with a learning rate of 1e-5. We train these metrics on the Daily Dialog dataset (Li et al. 2017) and choose models that have the lowest loss on the Daily Dialog evaluation data. We also test other pre-trained models, such as BERT (Devlin et al. 2019) or Roberta (Liu et al. 2019), but no performance improvement is observed.

Similar to Huang et al. (2020), the layer of graph attention networks (GATs) (Veličković et al. 2018) is 3, and the number of heads is set to 4, but we remove the contextualized encoding of context-response pair for model simplification. The training of the model is consistent with Huang et al. (2020) except that we modify the learning rate to 1e-5.

Overall Comparisons

Comparison Setting. In this part, we will compare 1) Deep AM-FM (Zhang et al. 2021) (the baseline of task1) and top 5 teams; 2) our different submissions on the test datasets. Table 3 lists the Deep AM-FM and top 5 with the highest average Spearman correlation on the test datasets. In Table 4, we list the comparison results of our 5 submissions. The baseline method Deep AM-FM and our submissions are introduced as follows:

[2] http://convai.io/2017/data/

[3] The statistics information of development datasets is listed in Table 1.

Method	J-A	E-A	N-A	DT-A	DT-C	DT-G	DT-R	DP-A	DP-C	DP-G	DP-R	Avg
Deep AM-FM	5.09	32.29	16.49	18.23	8.63	**16.84**	26.21	21.04	14.22	19.08	24.11	18.38
Top 1 (**ours**)	11.66	41.44	**29.88**	**32.64**	17.23	8.96	**44.76**	**45.60**	**32.53**	**21.98**	**54.76**	**31.04**
Top 2	8.52	38.11	26.61	31.78	**17.89**	8.52	43.81	45.36	32.78	21.47	54.35	29.93
Top 3	**26.15**	**47.56**	19.89	27.66	15.52	2.81	38.31	41.81	30.49	18.08	49.92	28.93
Top 4	12.73	32.11	26.47	29.97	15.56	6.16	42.47	43.43	31.46	18.85	53.10	28.39
Top 5	16.42	43.60	27.05	30.75	12.62	7.54	41.86	39.86	22.95	17.42	47.14	27.93

Table 3: The Spearman correlation (%) of baseline Deep AM-FM and top 5 teams on the test datasets. The test dataset JSALT, ESL, NCM, DSTC10-Topical, DSTC10-Persona are abbreviated as J, E, N, DT, DP respectively. And the evaluation qualities appropriateness, content, grammar, relevance are abbreviated as A, C, G, R respectively. The Avg column lists the average score of corresponding method on the 11 evaluation qualities. **Bold** denotes the best result for the corresponding quality.

Method	Submission Rank	Avg
Our submission 1	2	29.96
Our submission 2	1	**31.04**
Our submission 3	4	29.77
Our submission 4	5	29.64
Our submission 5	6	29.61

Table 4: The average Spearman correlation (%) of our different submissions on the test datasets. Each participant has up to 5 submissions, and we list the rank of our submissions in all submissions. **Bold** denotes the best result in our submissions.

- Deep AM-FM. A DNN-based automatic metric that measures the evaluation quality of dialogue generation along two dimensions: 1) Adequacy Metric: The semantic similarity between dialogue context and response; 2) Fluency Metric: The syntactic quality of the sentence construction.
- Top1-5. Top1-5 refer to the top 5 teams with the highest average Spearman correlation on test datasets of task1. And the top1 is our team.
- Our submission 1-3. The metric scores are integrated using our proposed CRS score composition method. To simplify the computation of the CRS method and improve the generalization ability of our metric, we simply set d_{ij} to 1, 2, 3, respectively in our submission 1-3.
- Our submission 4-5. We compute a specific d_{ij} for each dialogue quality in each dataset. In submission 4, a SimCSE pre-trained model is used, while in submission 5, a BERT pre-trained model is used instead, avoiding that the SimCSE model does not work well on the test set.

Comparison Results. We compare our approach with Deep AM-FM and the top 5 teams in Table 3. The performance comparison of our different submissions is shown in Table 4. These results support the following statements:

- Our MME-CRS achieves the highest average Spearman correlation (1.11% higher than the second) on five test datasets, which demonstrates the effectiveness of our proposed metric MME-CRS.

Method	Avg
MME-CRS (**ours**)	**31.04**
w/o FM	30.34
w/o RM	29.48
w/o TCM	27.78
w/o EM	30.05
w/o SM	30.96
w/o RM+TCM	11.07

Table 5: MME-CRS ablation study. Our submission 2 is chosen in this ablation as the first row. FM, RM, TCM, EM, SM represent fluency metric, relevance metric, topic coherence metric, user engagement metric and specific metric respectively. "w/o FM" means that fluency metric will not be used in the score composition. **Bold** denotes the best result in the ablation.

- Our method ranks first in 6 out of 11 dialogue evaluation qualities, demonstrating that our proposed evaluation metrics have a higher correlation with human judgments than baseline and other teams.
- Our submissions 1-3, which fix d_{ij} to a constant number, perform better than submissions 4-5, which indicate a constant d generalizes better when migrating to the test datasets. Furthermore, setting d_{ij} to 2 achieves the best performance on test datasets.

Ablation Study

Comparison Setting. In the final evaluation period, participants must submit a single score for each dialogue, and the organizers will compute the correlation between human scores of each quality with submitted scores. In Table 5, we remove fluency metric (FM), relevance metric (RM), topic coherence metric (TCM), engagement metric (EM), and specific metric (SM), respectively, to explore the importance of different metrics in the submitted scores. We take the SM composed of three metrics as a whole part to explore the influence of the SM. Considering that RM and TCM both rely on dialogue context, we also remove them together in our experiments.

Comparison Results. The comparison results of the ablation experiment are shown in Table 5. These results support

Method	J-A	E-A	N-A	DT-A	DT-C	DT-G	DT-R	DP-A	DP-C	DP-G	DP-R	Avg
MME-CRS	**11.66**	**41.44**	**29.88**	**32.64**	**17.23**	**8.96**	**44.76**	**45.60**	**32.53**	**21.98**	**54.76**	**31.04**
MME-Avg	8.23	37.19	28.88	30.64	13.42	6.67	41.60	42.83	26.59	19.83	47.21	27.55

Table 6: The comparison results of MME-CRS and MME-Avg. MME-Avg assigns equal weights to designed metrics, thus the composition score are simply the average of metric scores.

the following statements:

- TCM, RM, and EM contribute most to the performance. When we delete them from score composition, the final average Spearman correlation will drop 3.26%, 1.56%, and 1.01%, respectively.
- Coarse-grained RM and fine-grained TCM are a beneficial complement to each other. If we ignore one of them, the performance will drop slightly. However, if both of them are ignored, the average correlation will drastically drop to 11.07%.
- The improvement of SM can be ignored on the test datasets. We observe that many responses in test datasets tend to be very specific but are not relevant to the dialogue context. We infer that these models used to generate the response are overfitted on the test dataset.

The Effectiveness of CRS Method

Comparison Setting. Score composition is a significant component of open-domain dialogue because the full evaluation of dialogue usually depends on many aspects. In this part, we will compare the performance of MME-Avg, which simply averages the scores from different metrics, and our MME-CRS method.

Comparison Results. The comparison result is listed in Table 6, and the following statements can be drawn from the results.

- The average Spearman correlation score is significantly superior to that of MME-Avg (3.49% higher), indicating that our proposed CRS method can effectively integrate different scores to comprehensively measure the quality of dialogue.
- The correlation of MME-CRS is higher on all evaluation qualities, demonstrating that each evaluation quality can benefit from the score composition method CRS.

Conclusion

In this paper, we propose an open-domain dialogue evaluation approach composed of 5 groups of metrics to fully measure the quality of model response. Further, we propose a novel metric composition method called CRS. CRS models the relationship between metrics and evaluation qualities to comprehensively integrate sub-metric scores for dialogue evaluation. Experimental results on test datasets show that our proposed MME-CRS achieves the best performance, showing that our metric correlates better with human judgments. Compared with baseline and other teams, our approach obtains superior performance and ranks 1st in the final evaluation.

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-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*, 65–72.

Chen, Z.; Sadoc, J.; D'Haro, L. F.; Banchs, R.; and Rudnicky, A. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint arXiv:2111.02110*.

Danescu-Niculescu-Mizil, C.; and Lee, L. 2011. Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 76–87.

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*, 4171–4186.

Galley, M.; Brockett, C.; Gao, X.; Gao, J.; and Dolan, B. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv preprint arXiv:2104.08821*.

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*, 82–89.

Ghazarian, S.; Weischedel, R.; Galstyan, A.; and Peng, N. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7789–7796.

Gupta, P.; Mehri, S.; Zhao, T.; Pavel, A.; Eskenazi, M.; and Bigham, J. P. 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*, 379–391.

Hori, C.; and Hori, T. 2017. End-to-end Conversation Modeling Track in DSTC6. *dialog*, 888(107,506): 2–000.

Huang, L.; Ye, Z.; Qin, J.; Lin, L.; and Liang, X. 2020. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9230–9240.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Lan, T.; Mao, X.-L.; Wei, W.; Gao, X.; and Huang, H. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Transactions on Information Systems (TOIS)*, 39(1): 1–37.

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 Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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*, 225–235.

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*, 681–707.

Merdivan, E.; Singh, D.; Hanke, S.; Kropf, J.; Holzinger, A.; and Geist, M. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3): 762.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nübel, R. 1997. End-to-End evaluation in VERBMOBIL I. *Proceedings of MT Summit VI*, 232–239.

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*, 311–318.

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*, 4164–4178.

Rus, V.; and Lintean, M. 2012. An optimal assessment of natural language student input using word-to-word similarity metrics. In *International Conference on Intelligent Tutoring Systems*, 675–676. Springer.

See, A.; Roller, S.; Kiela, D.; and Weston, J. 2019. What makes a good conversation? How controllable attributes affect human judgments. 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)*, 1702–1723.

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*, 2430–2441.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, W. 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*, 196–205.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Tao, C.; Mou, L.; Zhao, D.; and Yan, R. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Weston, J.; Dinan, E.; and Miller, A. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, 87–92.

Wieting, J.; Bansal, M.; Gimpel, K.; and Livescu, K. 2016. Towards universal paraphrastic sentence embeddings. In *4th International Conference on Learning Representations*.

Yao, L.; Zhang, Y.; Feng, Y.; Zhao, D.; and Yan, R. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2190–2199.

Zhang, C.; D'Haro, L. F.; Banchs, R. E.; Friedrichs, T.; and Li, H. 2021. Deep AM-FM: Toolkit for automatic dialogue evaluation. In *Conversational Dialogue Systems for the Next Decade*, 53–69. Springer.

Zhang, R.; Guo, J.; Fan, Y.; Lan, Y.; Xu, J.; and Cheng, X. 2018. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1108–1117.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Zhao, T.; Lala, D.; and Kawahara, T. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 26–33.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 654–664.