C-PMI: Conditional Pointwise Mutual Information for Turn-level Dialogue Evaluation

Liliang Ren[*], Mankeerat Sidhu[*], Qi Zeng, Revanth Gangi Reddy, Heng Ji, ChengXiang Zhai
University of Illinois Urbana-Champaign
{liliang3, mssidhu2, qizeng2, revanth3, hengji, czhai}@illinois.edu

Abstract

Existing reference-free turn-level evaluation metrics for chatbots inadequately capture the interaction between the user and the system. Consequently, they often correlate poorly with human evaluations. To address this issue, we propose a novel model-agnostic approach that leverages Conditional Pointwise Mutual Information (C-PMI) to measure the turn-level interaction between the system and the user based on a given evaluation dimension. Experimental results on the widely used FED dialogue evaluation dataset demonstrate that our approach significantly improves the correlation with human judgment compared with existing evaluation systems. By replacing the negative log-likelihood-based scorer with our proposed C-PMI scorer, we achieve a relative 60.5% higher Spearman correlation on average for the FED evaluation metric. Our code is publicly available at https://github.com/renll/C-PMI.

1 Introduction

Evaluating dialogues is a multi-faceted task that demands consideration of diverse dimensions, which distinguishes it from the evaluation of task-oriented dialogue systems. Traditional n-gram-based evaluation metrics, such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), demonstrate weak correlation with human-annotated judgments due to the broad spectrum of potential responses in dialogues. As a result, researchers often resort to human evaluations to ascertain the quality and effectiveness of their generated system responses, especially for knowledge-guided dialog systems (Li et al., 2022; Fung et al., 2023; Lai et al., 2023).

Substantial research has been conducted on automatic evaluation metrics for dialogue (Yeh et al., 2021). These metrics can be classified into reference-based and reference-free categories. Reference-based metrics, which depend on comparing the system response to a human-written reference response, are generally inadequate for dialogue evaluation due to the inherent one-to-many nature of dialogues. The reference-free metric instead uses a computational model to generate a score for the system response with a given context.

Early models predominantly focus on a limited set of general features of dialogue generation quality, such as context coherency and fluency. Subsequent evaluation metrics investigated additional dimensions, such as USL-H (Phy et al., 2020), which combines relevance evaluation with fact-to-response selection. Holistic-eval (Pang et al., 2020) assesses content coherence, language fluency, self-consistency, and semantic appropriateness. D-Score (Zhang et al., 2021b) and Predictive Engage (Ghazarian et al., 2020) introduce response diversity and engagement scores. The recent FED (Mehri and Eskenazi, 2020a) metric encompasses 18 turn-level and dialogue-level metrics, including interestingness, likeability, and response flexibility. However, all of these methods do not model the interaction between the turn-level response and the dialogue history and regard them as an integrated context for score calculation.

In this paper, we focus on directly modeling user-system interactions through the lens of Mutual Information (Shannon, 1948; Ghassami and Kiyavash, 2017) and propose a novel scorer based on Conditional Pointwise Mutual Information (C-PMI), which effectively captures the turn-level interactions between the system and user with respect to a given hypothesis. We demonstrate that our approach results in a reference-free, training-free, automatic turn-level dialogue evaluation that significantly outperforms state-of-the-art methods with a comparable number of model parameters. Our contributions in this work are three-fold:

- A novel dialogue evaluation metric based on Conditional Pointwise Mutual Information (C-PMI) that effectively captures turn-level in-

[*] Equal contribution.

teractions between the system and user with respect to a given hypothesis.

- An unreferenced, training-free, automatic turn-level dialogue evaluation that significantly outperforms state-of-the-art methods with a comparable number of model parameters.

- A model-agnostic approach that can be served as a generalized alternative to the Negative Log-Likelihood (NLL) based evaluation metrics when interactions between previous turns need to be considered.

2 Related Work

Developing automatic evaluation metrics for dialog is challenging for several reasons: 1) Dialogues often have a one-to-many nature, rendering word-overlap metrics ineffective. To address this issue, metrics should be designed to be reference-free. 2) Given the limitless nature of conversation topics in open-domain dialogues, the dialogue evaluation metrics are expected to understand the semantic meaning of both the dialogue context and the generated responses. This necessitates a metric that can leverage pre-trained large language models and self-supervised training objectives. 3) Training dialogue evaluation metrics solely on labeled data can significantly restrict the metric's range, risking over-fitting to the training data in terms of conversation topics and response generation models. As such, recent metrics have started to incorporate self-supervised training objectives designed to capture various aspects of a dialogue, such as relevance, fluency, and interestingness among others.

Given the aforementioned challenges, large language models have become an integral part of dialogue evaluation. DialogRPT (Gao et al., 2020) employs an extended GPT-2 model trained on 147 million conversation-like interactions from Reddit. USR (Mehri and Eskenazi, 2020b) is an unsupervised, reference-free tool that takes advantage of the RoBERTa (Liu et al., 2019) model. USR employs a dialogue retrieval metric for assessing dialogue, where the metric is trained to differentiate between a ground truth response and a randomly sampled response. The FED metric (Mehri and Eskenazi, 2020a) utilizes DialoGPT (Zhang et al., 2020) due to its capacity for capturing knowledge, specifically within the context of conversations. It ignores the interaction between the user and the system and consider the dialogue history and the system response as an integral context, while our method explicitly captures such interaction through conditional mutual information.

3 Background

FED (Mehri and Eskenazi, 2020a) measures eighteen fine-grained qualities of dialogue without requiring comparison to a reference response or training data with ground-truth human ratings. The method leverages DialoGPT and uses the follow-up hypotheses as a means of evaluation, based on the assumption that the language model has learned to accurately measure the likelihood of the input sequence. Given a dialog context c, a system response r, and a scorer $\mathcal{L}$ that computes the average Negative Log-Likelihood (NLL) of a sequence with a language model θ, the predicted score for a pair of positive and negative hypotheses (p_i, n_i) is calculated as,

$$\sum_{i=1}^{|n|} \mathcal{L}\left(\{c, r, n_i\}, \theta\right) - \sum_{i=1}^{|p|} \mathcal{L}\left(\{c, r, p_i\}, \theta\right),$$

where $\{a, b\}$ means text b is appended to text a,and for each of the evaluation dimensions, $|p|$ and $|n|$ number of positive and negative hypothetical sentences are respectively pre-defined and used for reducing evaluation variance. For example, given a combined history $\{c, r\}$, the response is regarded as more interesting if the probability of DialoGPT generating a positive hypothesis (e.g., "That's really interesting!") is greater than the probability of it generating a negative one (e.g., "That's really boring.").

4 Conditional Pointwise Mutual Information based Turn-level Metric

For each of the dialogue turn t, our Pointwise Mutual Information (PMI) based metric is considering the dependencies between the following three random variables: the full dialogue history $\mathbf{r}_t = \{u_0, x_0, u_1, x_1, ..., u_t\} \sim R$ (where u_t is the user utterance), the system response $x_t \sim X$ and a hypothesis $h \sim H$. Ideally, we want to know how much correlation between the dialogue history and the system response causes the hypothesis to be a plausible entailment of the combined history, $\{\mathbf{r}_t, x_t\}$. We measure such correlation by calculating the Conditional Mutual Information (CMI) between the response and the history with a given hypothesis, *i.e.*,

$$I(R, X|H) = \mathbb{E}_{R,X,H}[\log \frac{p(\mathbf{r}_t, x_t|h)}{p(\mathbf{r}_t|h)p(x_t|h)}]$$
$$= \mathbb{E}_{R,X,H}[\log \frac{p(\mathbf{r}_t, x_t, h)p(h)}{p(\mathbf{r}_t, h)p(x_t, h)}].$$

Intuitively, if $I(R, X|H)$ is large, the hypothesis is less likely to be caused by the interaction (*i.e.*, the shared information) between R and X.

Since sampling the history on a turn-by-turn basis needs exponentially increasing computation, an accurate estimation of the CMI between these random variables is intractable. Therefore, we propose to measure the CMI by calculating the pointwise mutual information contained between the observed dialogue history and the system response when the hypothesis is appended to the combined history. Formally, we define our Conditional PMI (C-PMI) score between the observed dialogue history, the system response, and the hypothesis as follows,

$$\text{C-PMI}(\mathbf{r}_t, x_t|h) = \log \frac{p(\mathbf{r}_t, x_t, h)p(h)}{p(\mathbf{r}_t, h)p(x_t, h)}.$$

In practice, we estimate the probability of each sequence using the averaged Log-Likelihood (LL) obtained from a language model P_θ, *i.e.*,

$$\text{LL}(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n} \log P_\theta(s_i|\mathbf{s}_{<i}),$$

and our score is then computed as,

$$\text{C-PMI}(\mathbf{r}_t, x_t|h) = \text{LL}(\mathbf{r}_t, x_t, h) + \text{LL}(h)$$
$$- \text{LL}(\mathbf{r}_t, h) - \text{LL}(x_t, h),$$

which can be efficiently implemented using the modern deep learning framework. To retain the symmetric property of the mutual information, we also define a symmetric version of our score, C-PMI-SYM, by interchanging the response and the dialogue history, *i.e.*,

$$\text{C-PMI-SYM}(\mathbf{r}_t, x_t|h) = \frac{1}{2}(\text{C-PMI}(\mathbf{r}_t, x_t|h)$$
$$+ \text{C-PMI}(x_t, \mathbf{r}_t|h)).$$

For integrating our scorer with the existing evaluation system such as FED, we simply replace its NLL scoring function with our C-PMI scorer, and follow the original pipeline to get the final score for each of the data samples.

5 Experiments

5.1 Dataset

We evaluate our model on the turn-level annotated subset of the FED (Mehri and Eskenazi, 2020a) dataset. This subset consists of 455 data samples, each of which includes a dialog context, a system response, and eight human-annotated turn-level labels: Interesting, Fluent, Engaging, Specific, Relevant, Correct, Appropriate, and Understandable. The annotations are obtained through a survey with the options of No, Somewhat, Yes, or N/A. An additional overall impression label is measured using a five-point Likert Scale. The FED dataset is proposed to evaluate metrics as it is annotated with human quality judgments with conversations from Meena and Mitsuku bots (Adiwardana et al., 2020).

5.2 Baseline Metrics

We primarily compare our proposed reference-free and unsupervised metric with FED, but other baselines are also included as follows.

BARTScore (Yuan et al., 2021) is a text-scoring model based on BART (Lewis et al., 2020) and does not requiring any fine-tuning. BARTScore calculates the weighted log probability of text $\mathbf{y}$ given text $\mathbf{x}$:

$$\text{BARTSCORE} = \sum_{t=1}^{m} \omega_t \log P_\theta(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{x}),$$

where the weighted sum of the log probability of one text $\mathbf{y}$ given the other text $\mathbf{x}$ is used for scoring.

DynaEval (Zhang et al., 2021a) is an automatic evaluation framework for dialogue response generation tasks, designed to evaluate both turn-level and dialogue-level. The framework utilizes structured graph representations of dialogues and is trained on datasets that contain ground-truth human ratings.

5.3 Implementation Details

We follow the data pre-processing procedure as used by Yeh et al. (2021) for the FED dataset , and modify the scorer function as in the original FED repository. Following Yeh et al. (2021), we use a special "<|endoftext|>" token to connect each turn of the system responses and the user utterances for constructing a full sequence. The sequence is then fed into the *DialoGPT-large* language model to obtain the log-likelihood for calculating both the FED score and our C-PMI score.

Metrics	**Interesting**	**Fluent**	**Engaging**	**Specific**	**Relevant**	**Correct**	**Appro.**	**Und.**	**Avg.**
Supervised with Human Evaluations									
DynaEval	32.7	17.1	30.0	34.6	26.3	24.2	20.2	20.0	25.6
Unsupervised									
BARTSCORE	15.9	14.0	22.6	8.3	11.9	7.6	10.0	**12.0**	12.8
FED	32.4	-13.4	24.0	14.1	**19.9**	**26.2**	-9.4	1.3	11.9
FED*	32.5	*1.5*	17.6	23.0	13.4	15.9	*7.7*	*6.0*	14.7
FED + C-PMI-SYM	**48.2**	16.0	36.3	27.9	11.4	15.4	**17.8**	*9.8*	22.8
FED + C-PMI	**48.2**	**16.4**	**36.4**	**28.8**	13.5	17.4	**17.8**	*10.0*	**23.6**

Table 1: The Spearman correlations with human judgment on the FED Turn-level dataset. Italicized values indicate that they are not statistically significant (p > 0.05). We include the results from the supervised metric to showcase the power of our method. For the unsupervised metrics, the highest correlation is shown in bold and the second highest is underlined. * indicates our reimplementation. The results for DynaEval, BARTSCORE, and FED are from Fu et al. (2023). Appro. and Und. are respectively the abbreviations of the evaluation dimensions: Semantically Appropriate and Understandable.

6 Results & Analysis

Table 1 shows that our proposed metrics, FED+C-PMI-SYM and FED+C-PMI, outperform other methods in most of the evaluation dimensions, and is comparable to DynaEval which requires training on the evaluation dataset. Both FED+C-PMI-SYM and FED+C-PMI show substantial improvements in Interesting, Engaging, Specific, Semantically Appropriate and the Understable dimensions compared to our re-implemented FED metric. Notably, our metric even substantially outperforms DynaEval on the Interesting and the Engaging dimensions which conceptually needs an accurate measure of the interaction between the user and the system. This demonstrates the effectiveness of our approach in capturing turn-level interactions.

The performance of FED+C-PMI-SYM and FED+C-PMI is quite similar across most dimensions. However, FED+C-PMI shows slightly better performance in the Relevant, Correct, and Understandable dimensions, suggesting that the asymmetrical variant of the C-PMI calculation might provide more accurate evaluation scores in certain cases. We suspect that this is because interchanging the positions of the response and the dialogue history results in unnatural dialogue, which leads to worse probability estimation from the language models.

The results indicate that the proposed C-PMI-based turn-level metrics are capable of providing a more accurate evaluation of dialogue system responses compared to existing state-of-the-art methods. Moreover, our metric is unreferenced and training-free, which makes it particularly suitable for practical applications, such as responses selection and re-ranking.

7 Conclusion

In this paper, we introduce a novel dialogue evaluation metric based on Conditional Pointwise Mutual Information (C-PMI) that captures turn-level interactions between the system and user across various evaluation dimensions. The proposed metric is reference-free and training-free, outperforming state-of-the-art methods with a comparable number of model parameters. For turn-level dialogue evaluations, our experimental results demonstrate that this metric can serve as a generalized alternative to the Negative Log-Likelihood scorer for multi-dimensional evaluation metrics. We plan to extend our approach to other dialogue evaluation methods and explore its applicability to general text generation problems. We are also interested to see if our measure can improve the factual consistency evaluation for document-grounded dialogue or conversational question answering. Additionally, we will investigate incorporating our C-PMI-based metric into the fine-tuning process of LLMs.

Limitations

While our proposed method demonstrates promising results and outperforms several state-of-the-art techniques, it is important to acknowledge certain limitations.

- **Dependence on pre-trained LLMs:** Our method relies heavily on the pre-trained LLM's quality and the knowledge it has captured. As a result, any biases, inaccuracies, or limitations present in the LLM may directly impact the performance of our evaluation metric.

- **Lack of diversity in the dataset:** The FED dataset, which we use for evaluation, is primarily derived from conversations with the Meena and Mitsuku chatbots. Consequently, it is possible that our evaluation might not have better correlation with human ratings for other dialogue systems or more diverse conversational contexts.

- **Adaptability to new evaluation dimensions:** Our method currently focuses on eight turn-level metrics. Extending the method to incorporate additional or novel evaluation dimensions might require further investigation and calibration.

- **Computational cost:** The current implementation of our approach is around twice as slow as the baseline NLL-based method due to multiple times of the inferences of the language model. The efficiency of the implementation can be improved in the future by re-using the log-likelihood of the dialogue history.

- **Subjectivity in human judgments:** Our evaluation metric's correlation with human judgments serves as a key performance indicator. However, human judgments are inherently subjective, which could lead to inconsistencies or discrepancies in the evaluation results.

Despite these limitations, our proposed method presents a significant step forward in dialogue evaluation, offering a model-agnostic, unreferenced, and training-free approach that captures the human and the system interaction. Future work could address these limitations and explore additional dimensions of evaluation, further refining the method and its applicability across a broader range of dialogue systems and text evaluation systems.

Ethics Statement

In this study, we recognize the importance of ethical considerations in natural language processing and dialogue systems research. Acknowledging the potential biases in pre-trained LLMs and human judgments, we advocate for future research to investigate and mitigate these biases in evaluation metrics. We strive for fairness and inclusivity by designing our method to be generalizable and adaptable to various settings. As researchers, we are committed to responsible AI development and contribute to the ongoing discourse on evaluating dialogue systems, enabling the creation of more effective and ethical AI-powered conversational agents. We encourage the research community to continue discussing ethical considerations and promoting transparency in the field.

Acknowledgements

We would like to acknowledge support from the Amazon Alexa Prize as part of SocialBot Grand Challenge 5.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Yi Fung, Han Wang, Tong Wang, Ali Kebarighotbi, Prem Natarajan, Mohit Bansal, and Heng Ji. 2023. Deepmaven: Deep question answering on long-distance movie/tv show videos with multimedia knowledge extraction and synthesis. In *Proc. The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL2023)*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking-training with large-scale human feedback data. In *EMNLP*.

AmirEmad Ghassami and Negar Kiyavash. 2017. Interaction information for causal inference: The case of directed triangle. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1326–1330. IEEE.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 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, pages 7789–7796.

Tuan M. Lai, Giuseppe Castellucci, Saar Kuzi, Heng Ji, and Oleg Rokhlenko. 2023. External knowledge acquisition for end-to-end document-oriented dialog systems. In *Proc. The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL2023)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In *Proc. The 2022 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2022)*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

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

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

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

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

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. Dynaeval: Unifying turn and dialogue level evaluation. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Online.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021b. D-score: Holistic dialogue evaluation without reference. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2502–2516.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.