
Suvodip Dey and Maunendra Sankar Desarkar
Indian Institute of Technology Hyderabad, India
suvodip15@gmail.com, maunendra@cse.iith.ac.in

Abstract
In dialogue systems, automatically evaluating machine-generated responses is critical and challenging. Despite the tremendous progress in dialogue generation research, its evaluation heavily depends on human judgments. The standard word-overlapping based evaluation metrics are ineffective for dialogues. As a result, most of the recently proposed metrics are model-based and reference-free, which learn to score different aspects of a conversation. However, understanding each aspect requires a separate model, which makes them computationally expensive. To this end, we propose Dial-M, a Masking-based reference-free framework for Dialogue evaluation. The main idea is to mask the keywords of the current utterance and predict them, given the dialogue history and various conditions (like knowledge, persona, etc.), thereby making the evaluation framework simple and easily extensible for multiple datasets. Regardless of its simplicity, Dial-M achieves comparable performance to state-of-the-art metrics on several dialogue evaluation datasets. We also discuss the interpretability of our proposed metric along with error analysis.

1 Introduction
Dialogue systems research has seen massive advancements in recent years. It is not surprising to see models generating high-quality human-like meaningful responses nowadays. Despite this enormous progress, the evaluation of machine-generated dialogues remains a concern. Although many automatic metrics have been proposed, we still have to rely on human evaluation, which is tedious and costly. Thus, improving the quality of automatic dialogue evaluation is essential for the overall development of this evolving area.

The evaluation metrics for dialogue generation can be broadly divided into two classes: reference-based and reference-free. In reference-based metrics, the generated dialogue is evaluated with respect to one more reference utterance(s). The most popular reference-based metrics used in dialogue systems are standard word-overlapping based metrics like BLEU (Papineni et al., 2002), NIST (Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), Diversity (Li et al., 2016), and Entropy (Zhang et al., 2018b). However, these metrics have been shown to be ineffective because of the one-to-many nature of dialogues (Liu et al., 2016; Yeh et al., 2021). As a result, people started adopting learning-based referenced metrics like ADEM (Lowe et al., 2017), RUBER (Tao et al., 2017), BERT-RUBER (Ghazarian et al., 2019), PONE (Lan et al., 2020), BERTScore (Zhang* et al., 2020), BLEURT (Sellam et al., 2020), FBD (Xiang et al., 2021), Deep AM-FM (Zhang et al., 2021b), etc. However, reference-based metrics are not feasible for evaluation in an online setting where the reference response is unavailable. Also, collecting good-quality candidate responses is costly and requires human annotation. Hence, most of the recent efforts are being made in the direction of reference-free metrics.

In reference-free metrics, the generated dialogue is evaluated without any references. Here, most of the methods formulate the dialogue evaluation problem as one or more classification tasks and use the classification scores as the metric or sub-metrics. Metrics like Maude (Sinha et al., 2020) and DEB (Sai et al., 2020) learn to differentiate between correct and incorrect responses given the context. GRADE (Huang et al., 2020) and Dy-naEval (Zhang et al., 2021a) leverage graph-based methods, while DEAM (Ghazarian et al., 2022) relies on Abstract Meaning Representation (AMR) to evaluate dialogue coherence. MDD-Eval (Zhang et al., 2022) addresses the issue of multi-domain evaluation by introducing a teacher evaluator. The quality of a generated dialogue depends on multiple factors such as understandability, informativeness, coherence, etc. Metrics like USR (Mehri and Eskenazi, 2020b), USL-H (Phy et al., 2020), FED
(Mehri and Eskenazi, 2020a), HolisticEval (Pang et al., 2020), D-score (Zhang et al., 2021c), QualityAdapt (Mendonca et al., 2022) learn to compute various sub-metrics and then combine them to give a final score. For further improvement, \( FM^2 \) (Jiang et al., 2020) combines multiple metrics that are good at measuring different dialog qualities to generate an aggregate score. However, modeling different sub-metric requires a separate model or adapter, increasing the computational cost. Moreover, the decision boundary of the classification-based metrics depends on the quality of negative sampling (Lan et al., 2020), inducing training data bias.

In this work, we aim to address these issues by proposing **Dial-M**\(^1\), a Masking-based reference-free framework for Dialogue evaluation. The central idea of Dial-M is to mask the keywords of the current utterance and use the cross-entropy loss while predicting the masked keywords as the evaluation metric. Doing so avoids the requirement for multiple models and negative sampling, making the framework simple and easily extensible to multiple datasets. The keywords in the current utterance are obtained in an unsupervised manner. We show that Dial-M achieves comparable performance to various state-of-the-art metrics on several evaluation datasets, especially knowledge-grounded datasets like Topical-Chat. We observe that Dial-M can capture different aspects of a conversation. We also show that the Dial-M score can be interpreted by inspecting the masked words, which enables the scope for error analysis.

## 2 Dial-M Framework

Let \( D = \{ u_1, u_2, \ldots \} \) be a multi-turn conversation where \( u_i \) represents the utterance at turn \( i \). Let \( C = \{ c_1, c_2, \ldots \} \) be the set of conditions where \( c_i \) denotes the condition that is used to generate the \( u_i \). The condition can be knowledge, fact, persona, or other relevant information based on the task/dataset. The condition can be absent as well for conversations like chit-chat. For a given turn \( t \), the objective of dialogue generation is to generate \( u_t \) given \( D_{<t} \) i.e. \( \{ u_1, \ldots, u_{t-1} \} \) and \( C_t \) i.e. \( \{ c_1, \ldots, c_t \} \). The goal of the Dial-M framework is to learn a scoring function \( f: (D_{<t}, u_t, c_t) \rightarrow s \) where \( s \in \mathbb{R} \) denotes the quality of the generated response \( (u_t) \) given \( D_{<t}, u_t \) and \( c_t \) (if available). The details of our proposed framework are described as follows.

\(^1\)Code is available at github.com/SuvodipDey/Dial-M

![Figure 1: Dial-M Finetuning task.](image)

### 2.1 Pre-Training

We pre-train the RoBERTa (Liu et al., 2020) model with Masked Language Modeling (MLM) task on various conversational datasets. For a given conversation, the utterances are concatenated with a special token (eou). We consider only dialogue history for this MLM task, i.e., fact, persona, or any other conditions are ignored. We use RoBERTa-base\(^2\) with Language Model (LM) head as our base model. The masking probability is set to 0.15.

### 2.2 Finetuning

As discussed earlier, state-of-the-art evaluation metrics depend on multiple models to compute the final evaluation score. The main motivation for this work is to develop a lightweight alternative that can be trained using a single model and avoids negative sampling. To achieve this goal, we use a keyword masking task to finetune the pre-trained RoBERTa model (as shown in Fig. 1). For a given turn \( t \), we construct the RoBERTa input as text pair \( (D_t, c_t) \) or simply \( D_t \) if the condition is absent. The utterances of \( D_t \) are concatenated with the special token eou. Let \( K_t \) be the set of keywords in the current utterance \( u_t \). Let \( \hat{u}_t \) be the representation of \( u_t \) after masking the tokens associated with \( K_t \). Then we formulate our denoising task as predicting the masked tokens of \( u_t \) given \( D_{<t}, \hat{u}_t \), and \( c_t \) (if available). We use YAKE! (Campos et al., 2018, 2020), an unsupervised feature-based keyword extraction algorithm, to find the keywords. Further detail regarding YAKE! is provided in Appendix A.1. While finetuning, we ignore the utterances with no keywords.

In previous works, the standard MLM task has been used as a proxy for fluency or likability (Mehri and Eskenazi, 2020b; Pang et al., 2020). In contrast, focusing on the keywords helps to capture other important aspects like understandability, naturalness, and informativeness, which we later justify using the results of Table 2. Moreover, formulating the problem as an MLM task and the

\(^2\)huggingface.co/roberta-base
Table 1: Result comparison on various datasets with top-3 scores highlighted in bold. P and S indicate Pearson and Spearman’s coefficients, respectively. All values are statistically significant to $p < 0.05$, unless marked by *.

<table>
<thead>
<tr>
<th>Row</th>
<th>Metric</th>
<th>USR-Topical</th>
<th>USR-Persona</th>
<th>PredictiveEngage</th>
<th>HolisticEval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BLEU-4 (Papineni et al., 2002)</td>
<td>0.216</td>
<td>0.296</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>METEOR (Banerjee and Lavie, 2005)</td>
<td>0.336</td>
<td>0.391</td>
<td>0.253</td>
<td>0.271</td>
</tr>
<tr>
<td>3</td>
<td>BERTScore (Zhang* et al., 2020)</td>
<td>0.298</td>
<td>0.325</td>
<td>0.152</td>
<td>0.122*</td>
</tr>
<tr>
<td>4</td>
<td>BERT-RUBER (Ghazarian et al., 2019)</td>
<td>0.342</td>
<td>0.348</td>
<td>0.266</td>
<td>0.248</td>
</tr>
<tr>
<td>5</td>
<td>MADE (Sinha et al., 2020)</td>
<td>0.044*</td>
<td>0.083*</td>
<td>0.345</td>
<td>0.298</td>
</tr>
<tr>
<td>6</td>
<td>DEB (Sai et al., 2020)</td>
<td>0.180</td>
<td>0.116</td>
<td>0.291</td>
<td>0.373</td>
</tr>
<tr>
<td>7</td>
<td>GRADE (Huang et al., 2020)</td>
<td>0.200</td>
<td>0.217</td>
<td>0.358</td>
<td>0.352</td>
</tr>
<tr>
<td>8</td>
<td>HolisticEval (Pang et al., 2020)</td>
<td>0.147</td>
<td>-0.123</td>
<td>0.087*</td>
<td>0.113*</td>
</tr>
<tr>
<td>9</td>
<td>USR (Mehri and Eskenazi, 2020b)</td>
<td>0.412</td>
<td>0.423</td>
<td>0.440</td>
<td>0.418</td>
</tr>
<tr>
<td>10</td>
<td>USL-H (Phy et al., 2020)</td>
<td>0.322</td>
<td>0.340</td>
<td>0.495</td>
<td>0.523</td>
</tr>
<tr>
<td>11</td>
<td>$I_{M^2}$-overall (Jiang et al., 2022)</td>
<td>0.462</td>
<td>0.461</td>
<td>0.438</td>
<td>0.431</td>
</tr>
<tr>
<td>12</td>
<td>Dial-M (ours)</td>
<td>-0.432</td>
<td>-0.463</td>
<td>-0.464</td>
<td>-0.486</td>
</tr>
<tr>
<td></td>
<td>Ablation Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>with Random Masking</td>
<td>-0.320</td>
<td>-0.316</td>
<td>-0.359</td>
<td>-0.345</td>
</tr>
<tr>
<td>14</td>
<td>w/o Pre-training</td>
<td>-0.391</td>
<td>-0.429</td>
<td>-0.443</td>
<td>-0.489</td>
</tr>
<tr>
<td>15</td>
<td>w/o Finetuning</td>
<td>-0.290</td>
<td>-0.282</td>
<td>-0.288</td>
<td>-0.258</td>
</tr>
<tr>
<td>16</td>
<td>w/o Pre-training and Finetuning</td>
<td>-0.248</td>
<td>-0.248</td>
<td>-0.154</td>
<td>-0.144</td>
</tr>
</tbody>
</table>

The inclusion of dialogue conditions provides the flexibility to extend the framework to different kinds of conversational datasets without any additional annotation. For example, if the output of database queries (like system-act annotation in MultiWOZ (Budzianowski et al., 2018)) is converted into a natural sentence and used as the condition, Dial-M can be utilized for task-oriented conversation.

2.3 Dial-M Metric

To evaluate a generated response $u_t$, we first extract the set of keywords ($K_t$) from $u_t$. For each keyword in $K_t$, we mask the associated tokens and compute the cross-entropy loss to predict them using the finetuned RoBERTa model. We use the mean of these cross-entropy losses as our evaluation score. Let $k_{t,j}$ be the $j$th keyword in $K_t$. Let $T_{t,j}$ be the set of tokens associated with the word $k_{t,j}$. Let $\hat{u}_{t,j}$ be the representation of $u_t$ after masking the tokens $T_{t,j}$. Then the evaluation score ($s$) of the Dial-M metric is defined as:

$$s = \frac{1}{|K_t|} \sum_{j=1}^{|K_t|} \sum_{y \in T_{t,j}} \frac{1}{|T_{t,j}|} \log p(y|D_{<t}, \hat{u}_{t,j}, c_t)$$

(1)

We use YAKE! to extract the keywords. Since YAKE! is unsupervised and feature-based, it may not find all the relevant keywords. Thus, we also consider the words tagged with specific parts-of-speech (POS) as keywords to increase coverage. If no keyword is found in $u_t$, we consider all words as keywords. We observed that the utterances with no keywords are generally short and generic responses. As we are using cross-entropy loss, a lower score denotes a better response quality and vice-versa.

3 Experimental Setup

We use DailyDialog (Li et al., 2017), Persona-Chat (Zhang et al., 2018a), Wizard-of-Wikipedia (Dinan et al., 2019), and Topical-Chat (Gopalakrishnan et al., 2019) for both pre-training and finetuning Dial-M. We show our results on USR (Mehri and Eskenazi, 2020b), PredictiveEngage (Ghazarian et al., 2020), and HolisticEval (Pang et al., 2020) datasets for dialogue evaluation. USR is based on Topical-Chat and Persona-Chat, while PredictiveEngage and HolisticEval are based on DailyDialog. We call the Topical-Chat and Persona-Chat datasets of USR as USR-Topical and USR-Persona, respectively. We use spaCy (Honnibal and Montani, 2017) POS tagger along with YAKE! to find the keywords during evaluation. We analyzed the POS tags of co-occurring words in response ($u_t$) knowledge ($c_t$) pair in Topical-Chat train data and selected the most frequent POS tags ($NN, NNP, NNS, JJ, CD, VB, VBN, VBD, VBG, RB, BVP, VBZ, NNPS,$ and $JJS$) for our purpose. The rest of the details are provided in Appendix A.2.

4 Result and Analysis

Table 1 compares Dial-M with different metrics on four dialogue evaluation datasets. In Dial-M, a lower score is better, resulting in a negative correlation with the human scores. In Table 1,
Table 2: Correlation with different sub-metrics.

we can first observe that Dial-M outperforms the reference-based metrics (Rows 1-4). Secondly, it achieves comparable performance to state-of-the-art reference-free metrics. Thirdly, Dial-M performs relatively better for knowledge-grounded dialogues (USR-Topical and USR-Persona) than chit-chat (PredictiveEngage and HolisticEval). This is because the keywords of the current utterance generally align with context and the selected knowledge, which may not be the case for chit-chat. Nevertheless, the correlation values of Dial-M are close to the top-3 metrics for the chit-chat datasets. Table 2 shows the correlation of Dial-M with different sub-metrics on the USR dataset. Dial-M maintains a moderate correlation with all the sub-metrics, which justifies the utility of keyword masking in capturing different aspects of a conversation.

Rows 13-16 of Table 1 shows the result of our ablation study. In Row 13, we randomly mask 15% words of \( w_k \) instead of having a principled approach of identifying keywords and masking them while finetuning. We can observe that random masking degrades the performance except for HolisticEval. A similar observation can be seen in Row 15, where we do not use any finetuning i.e., the evaluation score is computed using the pre-trained model (described in Section 2.1). This conflicting behavior on HolisticEval can be due to the random chit-chat conversations in the dataset. In Row 14, we do not pre-train RoBERTa on dialogue datasets, which reduces the performance and shows the importance of pre-training. Row 16 displays the result with no training i.e. the scores are computed using the base RoBERTa model, resulting in poor performance.

5 Discussion

In this section, we discuss the interpretability and error analysis of Dial-M scores. Table 3 shows an illustrative example of Dial-M evaluation on USR-Persona. Masked words are shown in bold italics. Good cases (Responses 1-3). We can observe that Dial-M has given a low score to Response 1 in comparison to Responses 2 and 3, which correlates with the human scores. The reason for this low score can be deduced by looking at the masked words of Response 1, which are related to both context and condition (persona). In Response 2, masked words like red and blue are out of context, resulting in a higher Dial-M score. The masked words of Response 3 are slightly out of context in comparison to Response 1, resulting in an average score that is reflected in the human scores as well. Let us now analyze Response 4, which can be treated as a bad case because Dial-M finds it superior even though it is not the best response. The possible reason for the lower human score of Response 4 than Response 1 is the usage of “i get up early everyday”, which is not mentioned in the persona. However, the phrase “i get up early” is very common. Since Dial-M is pre-trained on MLM task, the prediction of “early” given “i get up” becomes easy, resulting in the lowest score. This is how we can interpret and perform error analysis of the Dial-M scores by inspecting the masked words. We observed that Dial-M generally assigns a low score to short, generic, and frequently used sentences where the masked word can be easily predicted from its neighbors. We aim to address this issue in our future work.

6 Conclusion

In conclusion, we propose Dial-M, a masking-based reference-free framework for dialogue eval-
We mask the keywords of the current utterance and use the cross-entropy loss while predicting the masked keywords as the evaluation metric. Formulating the problem as a keyword masking task avoids the requirement for multiple models and negative sampling, making the framework simple and easily extensible to multiple datasets. Dial-M achieves comparable performance to state-of-the-art metrics on several dialogue evaluation datasets. We also show the utility of keyword masking in capturing various aspects of a conversation and discuss the interpretability and error analysis of Dial-M scores. We want to explore better keyword extraction strategies in future work. We also want to investigate better techniques to handle the cases where no keywords are detected in the current utterance.

References


Appendix

A.1 YAKE!

YAKE! (Campos et al., 2018, 2020) is a lightweight unsupervised method for automatic keyword extraction. It is a feature-based system for extracting keywords from single documents, which supports texts of different sizes, domains, or languages. YAKE! builds upon unsupervised textual features (like casing, word frequency, word position, etc.) to find the most important keywords of a text, making it applicable to documents written in many different languages without the need for external knowledge. Thus, YAKE! does not rely on dictionaries/thesauri and requires no training against any corpora. However, it performs well and significantly outperforms other unsupervised methods on texts of different sizes, languages, and domains.

A.2 Implementation Details

We implemented Dial-M using PyTorch and Huggingface (Wolf et al., 2020) libraries in Python 3.10. All the experiments are performed on two devices of Nvidia DGX server with 32GB of memory each. The number of parameters in our pre-trained and finetuned model is 125M, the same as the RoBERTa-base model. The whole vocabulary is considered while predicting the tokens for the MLM tasks (both pre-training and keyword masking). The pre-training MLM task is trained for 30 epochs with a batch size 64 on a single GPU. The finetuning task is trained for 10 epochs with a batch size of 96 on two GPUs. We used AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate 1e-5 for both the training. The best model is selected based on minimum validation loss. The results of the other evaluation metrics in Table 1 and Table 2 are taken from the following references - Yeh et al. (2021); Mehri and Eskenazi (2020b); Jiang et al. (2022).

Fig. 2 shows the parts of speech (POS) of the co-occurring words in the response and corresponding knowledge in Topical-Chat (Gopalakrishnan et al., 2019) training data. We use the most frequent POS tags (NN, NNP, NNS, JJ, CD, VB, VBN, VBD, VBG, RB, VBP, VBZ, NNPS, and JJS) to mask the keywords during evaluation.
Figure 2: POS analysis on Topical-Chat train data.