Dial-M: A Masking-based Framework for Dialogue Evaluation

Anonymous ACL submission

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 submetrics. 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 DynaEval (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, IM^2 (Jiang et al., 2022) 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**, a **M**asking-based reference-free framework for **Dial**ogue 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. 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 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 $D = \{u_1, u_2, ...\}$ be a multi-turn conversation where u_i represents the utterance at turn i. Let $C = \{c_1, c_2, ...\}$ 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. $\{u_1, ..., u_{t-1}\}$ and C_t i.e. $\{c_1, ..., c_t\}$. 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.

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 condition is ignored. We use RoBERTa-base[1] with Language Model (LM) head as our base model. The masking probability is set to 0.15.

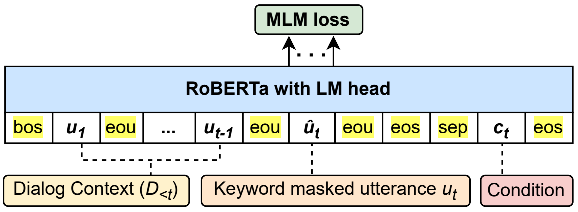

Figure 1: Dial-M Finetuning task.

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 keyword masking 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., 2020), an unsupervised feature-based keyword extraction algorithm, to find the keywords. While finetuning, we ignored 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 inclusion of dialogue conditions provide 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

[1] huggingface.co/roberta-base

Row	Model	USR-Topical		USR-Persona		PredictiveEngage		HolisticEval	
		P	S	P	S	P	S	P	S
1	BLEU-4 (Papineni et al., 2002)	0.216	0.296	0.135	0.090*	-	-	-	-
2	METEOR (Banerjee and Lavie, 2005)	0.336	0.391	0.253	0.271	-	-	-	-
3	BERTScore (Zhang* et al., 2020)	0.298	0.325	0.152	0.122*	-	-	-	-
4	BERT-RUBER (Ghazarian et al., 2019)	0.342	0.348	0.266	0.248	-	-	-	-
5	MAUDE (Sinha et al., 2020)	0.044*	0.083*	0.345	0.298	0.104	0.060*	0.275	0.364
6	DEB (Sai et al., 2020)	0.180	0.116	0.291	0.373	0.516	0.580	0.584	**0.663**
7	GRADE (Huang et al., 2020)	0.200	0.217	0.358	0.352	**0.600**	**0.622**	**0.678**	**0.697**
8	HolisticEval (Pang et al., 2020)	-0.147	-0.123	0.087*	0.113*	0.368	0.365	**0.670**	**0.764**
9	USR (Mehri and Eskenazi, 2020b)	**0.412**	**0.423**	**0.440**	0.418	**0.582**	**0.640**	0.589	0.645
10	USL-H (Phy et al., 2020)	0.322	0.340	**0.495**	**0.523**	**0.688**	**0.699**	0.486	0.537
11	IM^2-overall (Jiang et al., 2022)	**0.462**	**0.461**	0.438	**0.431**	-	-	-	-
12	Dial-M (ours)	**-0.432**	**-0.463**	**-0.464**	**-0.486**	-0.570	-0.592	**-0.590**	-0.598
	Ablation Study								
13	with Random Masking in Finetuning	-0.320	-0.316	-0.359	-0.345	-0.549	-0.547	-0.607	-0.630
14	w/o Pre-training	-0.391	-0.429	-0.443	-0.489	-0.556	-0.586	-0.567	-0.583
15	w/o Finetuning	-0.290	-0.282	-0.288	-0.258	-0.550	-0.549	-0.592	-0.613
16	w/o Pre-training and Finetuning	-0.248	-0.248	-0.154	-0.144	-0.508	-0.535	-0.540	-0.552

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 *.

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,

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

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. USR is based on Topical-Chat and Persona-Chat, while PredictiveEngage and HolisticEval are based on DailyDialog. 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*, *VBP*, *VBZ*, *NNPS*, and *JJS*) for our purpose. The rest of the details are provided in Appendix A.

Sub-Metric	Model	USR-Topical		USR-Persona	
		P	S	P	S
Understandable	USR	0.29	0.32	0.12	0.13
	Dial-M	**-0.35**	**-0.40**	**-0.18**	**-0.14**
Natural	USR	0.28	0.30	0.19	0.24
	Dial-M	**-0.37**	**-0.40**	**-0.28**	**-0.28**
Maintains Context	USR	**0.42**	0.38	**0.61**	**0.53**
	Dial-M	-0.37	**-0.40**	-0.40	-0.39
Engaging	USR	**0.46**	**0.46**	0.03	0.02
	Dial-M	-0.43	-0.45	**-0.33**	**-0.34**
Uses Knowledge	USR	0.32	0.34	**0.40**	0.32
	Dial-M	**-0.35**	**-0.37**	-0.34	**-0.37**

Table 2: Correlation with sub-metrics on USR data.

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, we can first observe that Dial-M outperforms the reference-based metrics (Row 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.

Row 13-16 of Table 1 shows the result of our ablation study. In row 13, we randomly mask 15% words of u_t instead of keyword masking 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.

Context ($D_{<t}$)	"hey . where are you from ? i'm from a farm in Wisconsin", "i love ice cream what is your favorite ? mine is chocolate", "mine is mint chocolate chip"
Condition (c_t) (Persona)	my wife and kids are the best. my favorite ice cream flavor is chocolate. i've three children. i'm a plumber. i love going to the park with my three children and my wife.
Response 1	my three ***kids love mint chocolate chip*** !
Human Score	Overall score: [5, 5, 5], Average: 5.0
Dial-M Score	0.1399
Response 2	i ***like*** the ***color red*** . i ***like*** the ***color blue*** .
Human Score	Overall score: [1, 2, 2], Average: 1.67
Dial-M Score	4.3131
Response 3	i ***like chocolate chip cookies***
Human Score	Overall score: [3, 4, 4], Average: 3.67
Dial-M Score	2.4582
Response 4	i get up ***early everyday*** and ***eat ice cream***
Human Score	Overall score: [3, 4, 5], Average: 4.0
Dial-M Score	0.1034

Table 3: Illustrative example of Dial-M evaluation on USR-Persona. Masked words are shown in bold italics.

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 a USR-PersonaChat conversation. Let us first analyze the good cases (Responses 1-3). We can observe that Dial-M has given a low score to Response 1 in comparison to Response 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 connected 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 evaluation. 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 masking and fallback strategies in future work.

References

Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7789–7796.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 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)*, pages 9230–9240, Online. Association for Computational Linguistics.

Zhihua Jiang, Guanghui Ye, Dongning Rao, Di Wang, and Xin Miao. 2022. IM^2: an interpretable and multi-category integrated metric framework for automatic dialogue evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11091–11103, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1).

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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

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

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

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.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. 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.

Ananya B. Sai, Akash Kumar Mohankumar, Siddharth Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI Conference on Artificial Intelligence*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing dialogue systems with distribution distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2192–2198, Online. Association for Computational Linguistics.

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.

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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

Chen Zhang, Luis D'Haro, Thomas Friedrichs, and Haizhou Li. 2022. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:11657–11666.

Chen Zhang, Luis Fernando D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li. 2021b. *Deep*

AM-FM: Toolkit for Automatic Dialogue Evaluation, pages 53–69. Springer Singapore, Singapore.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021c. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2502–2516.

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

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

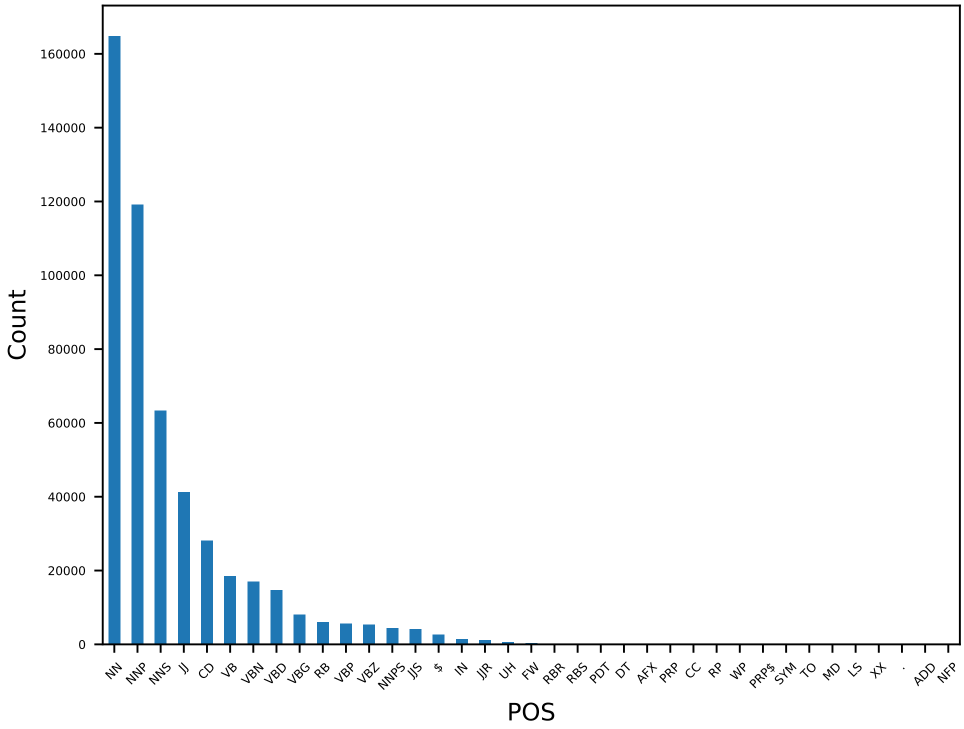

Figure 2: POS analysis on Topical-Chat train data.

A Appendix

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 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.