How to Choose How to Choose Your Chatbot: A *M*assively *M*ulti-*S*ystem *M*ulti*R*eference Data Set for Dialog Metric Evaluation

Huda Khayrallah‡ **Zuhaib Akhtar** **Edward Cohen**
Jyothir S V **João Sedoc**
hkhayrallah@microsoft.com {za203, jyothir}@nyu.edu jsedoc@stern.nyu.edu

Abstract

We release MMSMR, a **M**assively **M**ulti-**S**ystem **M**ulti**R**eference dataset to enable future work on dialog metrics and evaluation. Automatic metrics for dialogue evaluation should be robust proxies for human judgments; however, the verification of robustness is currently far from satisfactory. To quantify the robustness correlation and understand what is necessary in a test set, we create and release an 8-reference dialog dataset by extending single-reference evaluation sets and introduce a new language learning conversation dataset. We then train 1750 systems and evaluate them and publicly available large models on our novel test set and the DailyDialog dataset. In addition to the novel test set, we release model hyper parameters, inference outputs, and metric scores for each system on a variety of datasets.

Keywords: Metric Analysis, Dialog Metric Evaluation Set, Dialog System Evaluation Dataset

1. Introduction

Automatically evaluating social conversational agents (a.k.a. social dialogue systems or chatbots) is a challenging task that, if solved, would save time and money by making it easier to tune or evaluate such agents. There are three prevailing methods for evaluation: reference-based metrics $f(\hat{u}_t \mid \{r_t\})$, reference-free metrics $f(\hat{u}_t \mid u_{t-1} \ldots, u_0)$, and perplexity $f(\hat{u}_t)$, where $\hat{u}_t$ is the model generated response, $\{r_t\}$ are a set of references, and u_{t-1} is the previous utterance in the conversation. Evaluation metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and BARTScore (Yuan et al., 2021) are reported in the evaluation of open-domain chatbots models despite evidence of weak statistically significant correlation with human judgments (Liu et al., 2016; Yeh et al., 2021; Zhang et al., 2021). There is some evidence attributing the low correlation between reference-based metrics and human judgments to the "one-to-many" problem in conversational dialogue (Galley et al., 2015; Zhao et al., 2017; Gangal et al., 2021), whereby there can be multiple appropriate responses to a given input, and only a single 'ground-truth' reference response is used. Prior work demonstrated a higher correlation between automatic metrics and human judgments when using multiple references on the DailyDialog (Li et al., 2017) dataset (Gupta et al., 2019). Building upon this work, we extend the investigation to other datasets and employ a distinct methodology for gathering human annotations. A limitation of prior datasets is that the number of systems evaluated is extremely sparse (Zhang et al., 2021).

To address such limitations, we (will, on publication) release MMSMR, a **M**assively **M**ulti-**S**ystem **M**ulti**R**eference dataset. Our contributions are:

- We create and release a new conversational evaluation dataset based on hand-crafted conversations from material for teaching English as a second language[1] (ESL).[2]
- We collect and release multiple diverse 'ground-truth' human-generated reference responses for the ESL and NCM datasets.
- We train and release outputs of 1750 models on these data sets to enable the study of how metrics perform on a variety of models.
- We release the parameters to enable research on metrics without re-training new models.
- We release (sampled) inference from six large open-source models.
- We demonstrate the utility of the above contributions through analysis.

MMSMR is designed to test the robustness of dialog evaluation metrics in a statistically robust way.[3]

[1] rong-chang.com

[2] A subset of the prompts was made available online for use by other researchers in the past, but the dataset has not yet been published or released in full.

[3] As this paper provides a framework for evaluating the evaluation metrics, the title is intentional—we focus on choosing the method fro choosing.

2. Background & Related Work

Our work uses MMSMR to analyze automatic dialog metrics. We are far from the first to evaluate metrics using multiple annotations. Both multiple human-generate references, as well as multiple automatic references, have been explored (Gupta et al., 2019; Galley et al., 2015; Gangal et al., 2021). In particular, Gangal et al. demonstrate that automatically expanded reference sets improve correlations between human ratings and automatic metrics.

Other related prior work explores the relationships between metrics. In Yeh et al. (2021), 23 automatic evaluation metrics are evaluated on 10 datasets which are assessed to compare their shortcomings and strengths. In contrast to our work, past datasets rarely contained multiple references and also had few contrastive dialog systems. Similarly, Deriu et al. (2021) surveys new evaluation methods that reduce human interaction.

To the best of our knowledge large multi-system datasets do not exist for dialog evaluation; however, Zhang and Duh (2020) performed a grid search on machine translation and released it for research in hyper parameter optimization.

2.1. Dialog Evaluation Metrics

Automatic dialog evaluation metrics are mainly of two types: model based and rule based. The model based metrics measure the quality of responses that are generally trained. Rule-based metrics analyze the system response using heuristic rules based on human references and conversation context.

Several string overlap metrics are borrowed from other NLP tasks. In these metrics, the model output is compared to a human reference response. Bleu (Papineni et al., 2002), and Meteor (Banerjee and Lavie, 2005) come from machine translation, and Rouge (Lin, 2004) comes from summarization. Bleu is based on string matches using n-gram precision of the responses; Meteor includes synonyms and stems for computing the score. Rouge on the other hand uses n-gram recall. The effectiveness of these word overlap metrics has been a source of great debate (Liu et al., 2016; Lowe et al., 2017; Gupta et al., 2019; Galley et al., 2015).

Early model based metrics compute similarity between context and reference word embeddings (Mikolov et al., 2013b; Pennington et al., 2014; Mikolov et al., 2013a). BERTScore (Zhang et al., 2019) uses contextual embeddings for computing token similarity.

Prism (Thompson and Post, 2020) and BARTScore (Yuan et al., 2021) use sequence-level model scores. Prism uses a sequence-to-sequence paraphraser to score the output conditioned on human references, while BARTScore uses BART (Lewis et al., 2020), a denoising model. DialoRPT (Gao et al., 2020) is based on a set of GPT-2 models which are fine-tuned on a Reddit human feedback dataset.

USL-H (Phy et al., 2020) is a metric that is flexible to a task where a method is proposed to compound metrics called USL-H, which is Understandability, Sensibleness, and Likability in Hierarchy which is a single metric. USL-H combines three different models valid utterance prediction (VUP), next sentence prediction (NSP), and masked language model (MLM) where each model is trained on different tasks.

Prompt	**A:** I have a big surprise for you! **B:** Is it a new toy?
References	**R1)** No you have enough toys already. **R2)** I can't tell you yet! ... **R8)** Come out and look.
System Outputs	**S1)** It is a new car I got. I was surprised when I saw the price. I'm so very happy! **S2)** It is a toy that I got for my son. It is an electronic toy ...

Table 1: Dataset and Inference Snapshot on ESL2

3. Dataset Creation

Here we describe our collection methods for the 3500 multiturn conversations, multiple references for each prompt, and ratings for model generated responses.

3.1. ESL Multiturn Dataset

We scraped 3500 multiturn conversations (10+ turns) on a variety of topics that are used for instructing ESL speakers from `rong-chang.com`.[4] We randomly sampled 1000 snippets of 2 or 3 turns from the 3500 conversations. We name these the ESL2 and ESL3 test sets. A snapshot of ESL2 dataset including inferences is shown in Table 1.

3.2. Multireference Collection

In order to collect multiple references for each prompt, we created a HIT (human intelligence task) on Amazon's Mechanical Turk (AMT) and recruited crowdworkers. Each worker was shown 10 one-, two-, or three-turn conversations drawn from the Neural Conversational Model (NCM) (Vinyals and Le, 2015; Sedoc et al., 2019), ESL2, and ESL3

[4] We received permission to scrape and distribute a subset of the data.

datasets and asked to provide 2 to 5 responses to the last turn in each conversation.[5]

Beyond our quality control filtering, we analyzed the following: the average Jaccard distance of responses both for workers against themselves and against all of the provided responses for a prompt, the average number of responses provided by workers, and the fatigue factor for each of the prompt datasets. Across each of our datasets, the average Jaccard distance between each reference is high (at or near .9 across the board). Therefore, we conclude that there is **high diversity among the collected references**. This fact is key to the success of evaluation using multiple references (Gangal et al., 2021).

3.3. Model Responses

In order to understand how different metrics are able to distinguish between the quality of different models, we needed a large, diverse collection of model outputs. We collected these by using large pretrained models, and by training our own models.

The pretrained models we used are: Blenderbot (Shuster et al., 2022), Open-Assistant SFT-4 (Ope, 2023), Koala (Geng et al., 2023), MPT-7B-Chat (MPT, 2023), FastChat-T5 (Zheng et al., 2023), and Vicuna (Chiang et al., 2023). We employed a variety of temperature and sampling strategies for each model to generate diverse outputs.

Following Khayrallah and Sedoc (2020), we trained 1750 Transformer (Vaswani et al., 2017) chatbots in FAIRSEQ using base parameters from the FLORES benchmark for low-resource MT (Guzmán et al., 2019). In order to explore the full space of models with a variety of performance levels, we performed a hyperparameter sweep of regularization parameters, including SentencePiece (Kudo and Richardson, 2018) vocabulary size, dropout, attention & relu dropout, and label smoothing. We also used 8 different decoding strategies. We trained on the DailyDialog corpus (Li et al., 2017), as released by ParlAI (Miller et al., 2017).[6]

4. Methodology

To validate the utility of our dataset, we ask a few basic questions about the metrics. In particular, we aim to validate or challenge relationships between well-established metrics. Our approach is to evaluate outputs using multiple references rather than a single reference. For multiple models' responses to the same prompts, we use multiple evaluation metrics to score each of them.

We explore: (1) The Pearson and Spearman correlation between metric evaluations and human evaluations, (2) the Kendall rank correlation coefficient between metric evaluations and human evaluations, and (3) the relationship between output similarity and metric evaluations.

5. Analysis

Mathur et al. (2020) showed that correlating a machine translation metric with human judgments is far easier when considering all systems (including very weak ones) than when only considering top systems. Text simplification metrics also have similar behavior, where the correlation between metrics and human judgments decreases when filtered by system quality (Alva-Manchego et al., 2021).

This is somewhat intuitive: truly terrible systems are easier to differentiate from good ones. Therefore, we consider how well the metrics correlate overall, and when only considering the top systems.

We define top scoring as any system that is in the 99th percentile of systems on any metric. Figure 2 shows that top scoring systems constitute a large percentage of systems overall, which further highlights the disagreement between metrics. 48% of the systems are in the 90th percentile or above on some metric for NCM. If the metrics were in perfect agreement, only 10% of system would be in the 90th percentile. With so little agreement, it can be particularly hard to know which metrics to trust, highlighting the need for such a dataset for further metrics research. Figure 1 shows Spearman correlations between the various metrics (also see additional tables in the appendix). The bottom left half of each table shows the correlation between the metrics on all systems. The top right half shows the correlation between the top scoring systems.

Unsurprisingly, correlations are much stronger overall when comparing all systems rather than only comparing the top systems.

Meteor, Bartscore, USL-H and nup do not correlate well with other metrics, and have negative correlations in many settings. We note that Metor and Bartscore correlate well with each other. A lack of correlation with other metrics is not necessarily an indication of quality of a particular metric. But rather, this shows that there is poor agreement amongst metrics, and the metric chosen can have a large impact on the final ranking of ssystems.

6. Conclusion

We (will, upon publication) release MMSMR, a **M**assively **M**ulti-**S**ystem **M**ulti**R**eference dataset to

[5]The HIT html will be available in the supplemental materials. We use the AMT filters of location:US, approval rate > 95, approved HITs > 1000. Further details will be available upon publication.

[6]Information for replication and about hyperparameters will be available upon publication in the Appendix.

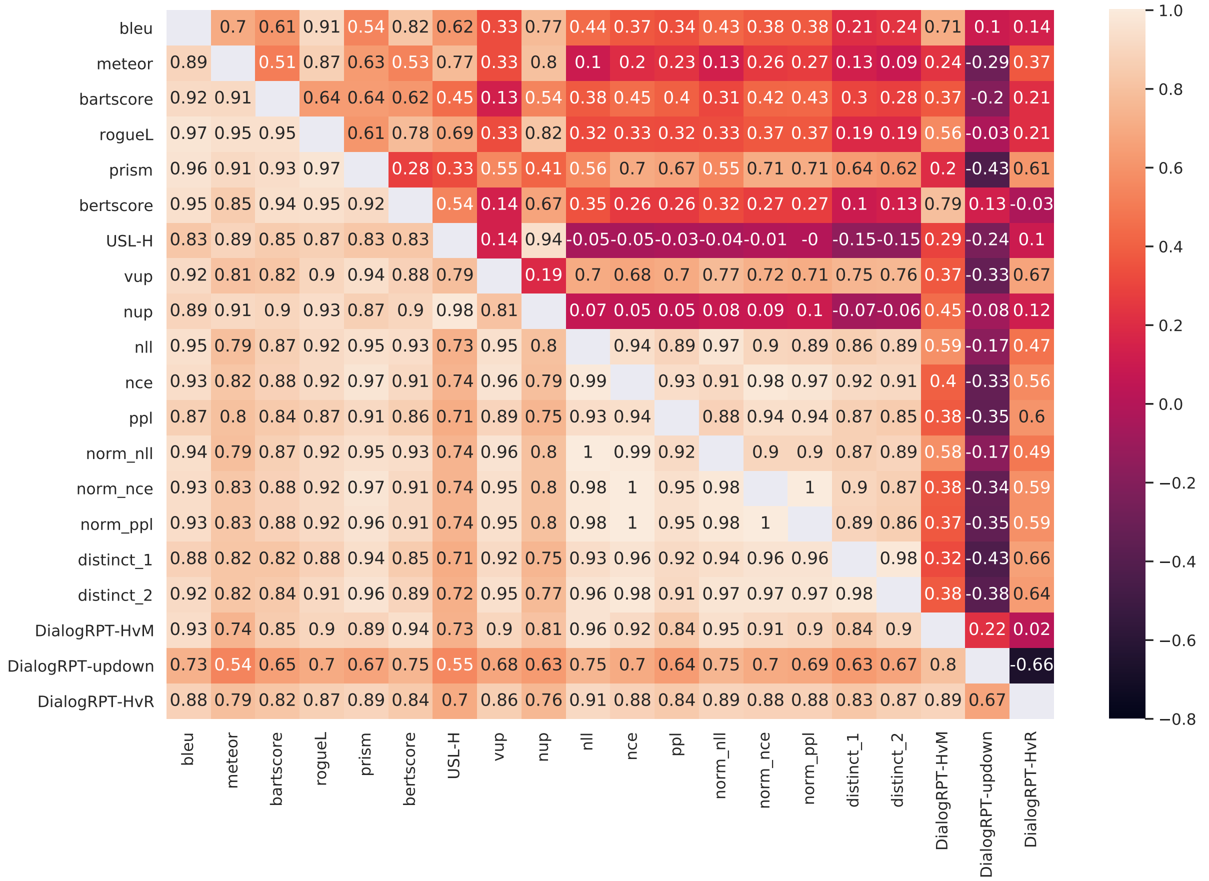

Figure 1: Spearman correlations between various metrics on the ESL3 test set. The bottom left includes all systems, the top right is the top ones.

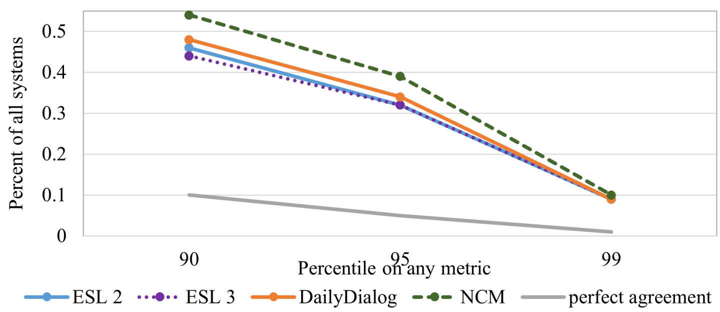

Figure 2: The percent of data retained when thresholding on a percentile for any of the metrics. The dotted grey line shows the percentage that would be retained if all metrics were in perfect agreement.

enable future work on metrics and evaluation for dialog. The dataset contains 1000 two and three-turn prompts with multiple human-generated references. We train 1750 systems and evaluate them on our novel test set and the DailyDialog dataset. We also evaluate publicly available models on these data sets. Our analysis of the metrics shows that the correlations are lower when considering only the top systems than when considering all systems. Our findings show the utility of this novel test set, and model hyper parameters, inference outputs, and metric scores for each system on a variety of datasets.

Limitations

We note that our contribution is meta-evaluation, not evaluation. Our goal is not to determine the best metric, but rather to provide a data set with which future work can try to answer that question.

We also only consider open source models, and therefore we do not use the recent OpenAI models.

This work's focus is only on English language datasets. More morphologically rich and/or lower resource languages may present additional challenges for evaluation. We hope this work motivates future work on meta-evaluation of chat-bots in more languages.

Our work focuses on next turn utterances in response to a prompt rather than at a dialog-level. This means that our dataset might not be as appropriate for examining dialog-level metrics.

Ethics Statement

Chatbots and their potential ethical impact have come into recent focus. This work focuses on the meta-evaluation of chit-chat bots, with a focus on quality. We note that there are other impacts of such models that must be considered before potential use of them. The decisions used in evaluating models determine what kinds of models will impact users.

2023. Introducing MPT-7B: A new standard for open-source, commercially usable LLMs. Blog post.

2023. Open Assistant. https://huggingface.co/OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5. Accessed: 2023-05-30.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China. Association for Computational Linguistics.

Varun Gangal, Harsh Jhamtani, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Improving automated evaluation of open domain dialog via diverse reference augmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4079–4090, Online. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *arXiv preprint arXiv:2009.06978*.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 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*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Huda Khayrallah and João Sedoc. 2020. SMRT chatbots: Improving non-task-oriented dialog with Simulated Multiple Reference Training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4489–4505, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

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.

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.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. 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.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. *arXiv preprint arXiv:2011.00483*.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*.

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

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint arXiv:2111.02110*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xuan Zhang and Kevin Duh. 2020. Reproducible and efficient benchmarks for hyperparameter optimization of neural machine translation systems. *Transactions of the Association for Computational Linguistics*, 8:393–408.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 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)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.