Improving Automated Evaluation of Open Domain Dialog via Diverse Reference Augmentation

Harsh Jhamtani * [1] **Varun Gangal** * [1] **Eduard Hovy** [1] **Taylor Berg-Kirkpatrick** [2]
[1] School of Computer Science, Carnegie Mellon University
[2] Computer Science and Engineering. University of California San Diego
{vgangal,jharsh,hovy}@cs.cmu.edu, tberg@ucsd.eng.edu

Abstract

Multiple different responses are often plausible for a given open domain dialog context. Prior work has shown the importance of having multiple valid reference responses for meaningful and robust automated evaluations. In such cases, common practice has been to collect more human written references. However, such collection can be expensive, time consuming, and not easily scalable. Instead, we propose a novel technique for *automatically* expanding a human generated reference to a set of candidate references. We fetch plausible references from knowledge sources, and adapt them so that they are more fluent in context of the dialog instance in question. More specifically, we use (1) a commonsense knowledge base to elicit a large number of plausible reactions given the dialog history (2) relevant instances retrieved from dialog corpus, using similar past as well as future contexts. We demonstrate that our automatically expanded reference sets lead to large improvements in correlations of automated metrics with human ratings of system outputs for DailyDialog dataset. [1]

1 Introduction

Evaluation by human annotators perhaps give the best insights into quality of machine generated natural language outputs. However, they can be expensive and time consuming. Much focus has therefore been on automated evaluation methods which correlate with human evaluations. Automated metrics such as BLEU (Papineni et al., 2002) work well for tasks such as machine translation, but often correlate poorly with human ratings in tasks such as open domain dialog which admit a wide variety of valid response for given context, often due to small number of human written references (Zhao et al., 2017; Sai et al., 2020b). Prior work (Sugiyama et al., 2019; Gupta et al., 2019) has demonstrated that having multiple valid references for the same context leads to automated metrics being better correlated to human judgements for appropriateness. However, collecting human written responses is difficult to scale, can be costly, and may find it difficult to cover a large variety of correct responses (Celikyilmaz et al., 2020).

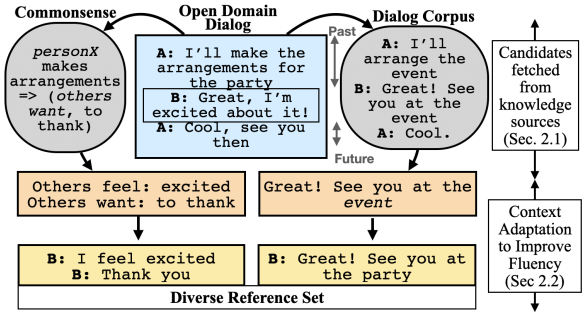

Figure 1: We propose automatic ways to collect references sans any crowd-sourcing, through two types of knowledge sources: commonsense and retrieved instance knowledge, followed by automated adaptation to make them more fluent in the target contexts.

In this work, we automatically extract a large number of diverse references to be used with such reference-based metrics, without resorting to expensive crowd-sourcing. Intuitively, since open-domain dialog pertains to everyday life, its utterance text tends to re-instantiate from a large but limited pool of situations (Schank, 1972) e.g friends debating politics etc, with variation only on some details e.g country discussed. Hence, knowledge encapsulating a wide scope of situations can serve as one starting point to automatically seed a set of diverse references. We first fetch plausible candidates from two types of knowledge sources (Figure 1). Such knowledge sources provide ready and easy access to a large number of potentially appropri-

* VG and HJ contributed equally for this paper. Order decided by coin flip.

[1] Code and data are available at https://github.com/harsh19/Diverse-Reference-Augmentation/

ate and diverse references. However, all retrieved instances may not be directly useful. As such, to achieve more fluent references, we propose techniques to adapt the candidate references based on the context (e.g change country being discussed). Note that since we are interested in creating references for only evaluating appropriateness of system outputs, our techniques can rely on broader data sources compared to dialog models. For example, we use future context and human written reference for retrieval, while a dialog model cannot.

Our contributions are as follows: (1) We propose a method for automated reference set augmentation for automated dialog evaluation. Compared to collecting more human-written responses, our approach is inexpensive and scalable, and fetches a diverse set of references. (2) We observe high correlations of various automated metrics with human ratings when proposed reference augmentation is applied to the test split of DailyDialog dataset (Li et al., 2017). We additionally observe that paraphrasing, a popular data augmentation technique, performs much worse. (3) We employ novel use of commonsense knowledge and dialog corpus instances, and unsupervised techniques for adapting retrieved references into more fluent forms.

2 Method

Figure 1 shows an overview of our proposed methodology. We first fetch plausible candidates from two types of knowledge sources. Thereafter, the retrieved candidate references are adapted so that they are fluent in the target context. We refer to our proposed method as **SCARCE** (**SC**alable **A**utomated **R**eference **C**onstruction for **E**valuation).

2.1 Knowledge Sources

Pre-trained Commonsense Model Much open domain dialog is based on everyday matters. We posit that extracting inferences about a situation using a commonsense knowledge base could be useful in identifying a wide variety of plausible reactions for a given dialog context. For example, a person making arrangements for an event might receive thanks from others (Figure 1). We utilize COMET (Bosselut et al., 2019) an off-the-shelf commonsense knowledge model built on ATOMIC (Sap et al., 2019a) or ConceptNet (Speer et al., 2017) corpus. It can be used to elicit commonsense inferences.

COMET-ATOMIC provides inferences on cause-effect interrelations between events pertaining to nine relation types such as *oReact* (effect on others due to the event), and *oWant* (inferences about wants of the receiver of event). Given an utterance from the previous speaker, we draw up to 5 inferences pertaining to each of oEffect, oReact, and oWant relation types to construct plausible references for the target response. For example, for an utterance 'I will make the arrangements. It will be great.', one of the inferences corresponding to oEffect is 'feel excited', depicting a plausible state of the next dialog speaker. However, such outputs are typically phrases, and we discuss transformation to fluent sentences in Section 2.2. Similarly, we use inferences pertaining to 'CausesDesire' and 'HasFirstSubevent' relation types from COMET-ConceptNet.

Dialog Corpus Retrieval For a test dialog context under consideration, one is likely to find similar contexts occurring in some of the training dialogues, given a sufficient number of them. Using retrieval, we can identify such contexts and use their responses as pseudo-references for the test-time response. Specifically, for retrieval, we use the BM25 function $S_{bm25}(x, y)$ (Robertson et al., 1995) to compare each element $\{d_t^{past}, d_t^{resp}, d_t^{future}\}$ of the turn under evaluation d_t (the query) with those of the candidate turn $x_{t'}$, $\{x_{t'}^{past}, x_{t'}^{resp}, x_{t'}^{future}\}$. Here, d_t^{past} and d_t^{future} are the windows of turn sequences before and after response d_t^{resp}.

Our approach is related to Galley et al. (2015), who propose Δ-BLEU measure which uses retrieval to produce pseudo-references. However, unlike here, they require annotator quality scores to weigh them during evaluation. Moreover, though we utilize retrieval for evaluation, methods of this kind have found success in many generation setups (Li et al., 2018; Peng et al., 2019; Khandelwal et al., 2019). Besides being automatic, our method differs from the above ones in that it explores the added utility of future information for retrieval. For instance, for the dialog shown in Figure 1, besides matching "Great!" in the response, our retrieval benefits from matching "cool" in the future.

2.2 Context Adaptation

We note that commonsense knowledge outputs are incomplete sentences, and we use simple templates to convert them to fluent sentences e.g. 'feels ex-

cited' gets transformed to 'i feel excited'. (Detailed templates in Appendix B) Further, we note that references from knowledge sources are often not fluent for the target context. For example, 'event' in the retrieved reference shown in Figure 1 can be updated to 'party' to construct a more apt reference. To adapt the retrieved text to better fit the target context we use employ an unsupervised decoding procedure, based on the approach of Qin et al. (2020), that uses gradient ascent to search for output text that maximizes (1) fluency with the left context (approximated by the likelihood of the output text under a pretrained GPT-2 model) and (2) similarity to the original text from the knowledge source (approximated by the likelihood of the original text under the output text's token-level word distributions). The method utilizes a heuristic update procedure to iteratively refine a differentiable proxy for the output text (a sequence token-level word distributions), while keeping the model parameters fixed. More details can be found in Qin et al. (2020) and in Appendix B.

3 Experiments

We investigate the extent to which automated metrics on an evaluation dataset correlate with human ratings of system outputs. We use the human ratings collected by Gupta et al. (2019), who collected utterance level human ratings using Amazon Mechanical Turk (AMT). They used a collection of 100 dialogue contexts that are randomly selected from the DailyDialog dataset. The generated response from various methods are rated in terms of appropriateness (from 1-5, with 5 denoting the best) by 5 different AMT workers. They collected and considered outputs from following methods: CVAE (Zhao et al., 2017), HRED (Serban et al., 2016), Seq2Seq (Vinyals and Le, 2015), Dual-encoder (Lowe et al., 2015), and Human-written responses. We report Spearman rank correlation (Spearman, 1961) and Kendall Tau rank correlation (Kendall, 1938) of human ratings against ngram overlap metrics such as BLEU (2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and embedding based metrics like cosine similarity of average word embedding (EmbeddingAvg) (Wieting et al., 2016) or Skip Thought Embedding (Kiros et al., 2015), and precision (BERT-Prec) and recall (BERT-Rec) components of BertScore (Zhang et al., 2020).

We compare the correlations across following setups: **SINGLE** (Li et al., 2017): Original DailyDialog dataset which had one reference per context; **SCARCE-SINGLE**: Proposed method along with SINGLE reference; **MULTI** (Gupta et al., 2019): 4 human written references. **SCARCE-MULTI**: Reference responses from the proposed method along with MULTI references. Additionally, we report the results when using PARAPHRASE instead of SCARCE: **PARAPHRASE-SINGLE** and **PARAPHRASE-MULTI**. Paraphrasing is a popular approach for automated data augmentation. Paraphrasing via backtranslation (BT) (Sennrich et al., 2016) is known to be an effective, domain-independent way to generate good quality paraphrases (Wieting and Gimpel, 2017). We use the BT model from (Xie et al., 2020) with its default hyperparameters to sample 5 paraphrases per human written reference

Results: We observe that most of the metrics show large improvements in correlations to human ratings for appropriateness when used along with SINGLE or MULTI (Table 1). In fact, rank correlations across most of the metrics are better for SCARCE-SINGLE compared to MULTI, even though former uses only single human written reference while latter uses upto 5 human written references[2]. Additionally, we observe that PARAPHRASE produces little or no improvements in correlations with human ratings (Table 1). We posit that for a given response, alternate responses constitute a strictly richer subspace than that of response paraphrases, which tend to be lexico-syntactically variant but semantically invariant.

Analyzing the impact of various components: To understand the impact of various components, we report Spearman rank correlation scores for BLEU-4 and BERT-Prec metrics with some variants of SCARCE-SINGLE (Table 2). We note that considering only one knowledge source (COMMONSENSE-only, RETRIEVAL-only) leads to good Spearman rank correlations of automated metrics to human ratings. Thus, the additive effect (SCARCE-SINGLE) shows rather small incremental benefit. Moreover, RETRIEVAL by itself does better than COMMONSENSE, though at smaller

[2]Rank correlations for SINGLE and MULTI deviate from the values in Gupta et al. (2019), who (in private communication with us), confirmed that the final dataset and code available on their repo does lead to the numbers we report.

Spearman Rank Correlation / Kendall Tau Rank Correlation						
Setup	1 human written reference			4 human written references		
Dataset	SINGLE (Li et al., 2017)	PARAPHRASE -SINGLE	SCARCE -SINGLE(Ours)	MULTI (Gupta et al., 2019)	PARAPHRASE -MULTI	SCARCE -MULTI(Ours)
BLEU-4	0.09 / 0.07	0.13 / 0.09	0.30 / 0.21	0.28 / 0.20	0.27 / 0.19	**0.36** / 0.25
BLEU-3	0.06 / 0.04	0.11 / 0.07	0.29 / 0.20	0.24 / 0.17	0.24 / 0.17	0.35 / 0.24
BLEU-2	0.04 / 0.03	0.08 / 0.06	0.28 / 0.19	0.20 / 0.14	0.21 / 0.14	0.33 / 0.23
BLEU-1	0.02 / 0.02	0.06 / 0.04	0.25 / 0.17	0.19 / 0.13	0.18 / 0.12	0.29 / 0.21
ROUGE-L	0.07 / 0.05	0.09 / 0.06	0.26 / 0.18	0.20 / 0.14	0.20 / 0.14	0.32 / 0.22
METEOR	0.11 / 0.07	0.09 / 0.06	0.24 / 0.17	0.23 / 0.16	0.22 / 0.15	0.30 / 0.21
EmbeddingAvg	0.03 / 0.02	0.02 / 0.01	0.02 / 0.02	0.10 / 0.07	0.10 / 0.07	0.08 / 0.05
SkipThought	-0.00 / 0.00	-0.03 / -0.02	0.09 / 0.07	0.07 / 0.05	0.05 / 0.04	0.13 / 0.10
BERT-Prec	0.27 / 0.19	0.28 / 0.19	0.38 / 0.26	0.32 / 0.22	0.32 / 0.22	**0.41** / 0.28
BERT-Rec	0.10 / 0.06	0.09 / 0.06	0.24 / 0.16	0.23 / 0.16	0.21 / 0.15	0.30 / 0.21
Max. value	0.27 / 0.19	0.28 / 0.19	0.38 / 0.26	0.32 / 0.22	0.32 / 0.22	0.41 / 0.28

Table 1: Utterance level Spearman Rank Correlation (1961) and Kendall Tau Rank Correlations (1938). (1) SCARCE-SINGLE augments the original single human written response (SINGLE) in DailyDialog dataset (Li et al., 2017) using proposed method. It leads to large improvements in correlations across most of the metrics, when compared to SINGLE. (2) SCARCE-MULTI augments the MULTI dataset, again leading to improvements in correlations to human ratings, especially for BLEU and BERT-Prec metrics.

Method	BLEU-4	BERT-Prec
SCARCE-SINGLE	0.30	0.38
SCARCE-SINGLE variants:		
COMMONSENSE only	0.24	0.31
RETRIEVAL only	0.29	0.36
RETRIEVAL only (5% corpus)	0.17	0.28
W/O CONTEXT-ADAPT	0.26	0.37

Table 2: Analyzing the impact of various components

corpus availability (e.g. 5%), COMMONSENSE performs better. Finally, not using context adaptation (W/O CONTEXT-ADAPT) leads to significant performance drop.

Qualitative Examples To illustrate our approach, we present a couple of examples in Table 3. (A wider selection of examples can be found in Appendix Table 6.)

Quality of Auto-generated References: We check the quality of SCARCE references by recruiting human annotators, showing them the reference along with the dialog context, and requesting them to tag each reference as appropriate, neutral, or not-appropriate, with respect to the dialog context. We randomly select 150 responses each from SCARCE and MULTI for this purpose. We observe that in about 29% of the references from SCARCE (fully automatically generated) were annotated as not appropriate, compared to 7% for MULTI, demonstrating fair quality of augmented responses from SCARCE (Additional details and results in Appendix). We do note that the ones marked as not relevant/appropriate can often be tweaked easily by a human to transform them into valid responses – demonstrating the possibility of exploring human-in-the-loop setups along with SCARCE to collect even better references.

4 Discussion

Transferability to more languages: Transferability of our approach to more languages is one aspect that merits discussion. While commonsense resources aren't readily available in all languages, a workaround can be to use off-the-shelf MT to translate before querying into English versions of the commonsense resources, and then translate back retrieved information. Furthermore, we note that while commonsense knowledge was useful, removing the COMMONSENSE method and relying on retrieval alone causes only relatively modest drop in performance (see Table 2). Thus, for languages lacking commonsense resources, one may still attain good gains in reference based evaluation by retrieving and adapting from dialog corpus alone.

Reference-less metrics: We note that while comparisons of using proposed approach against using reference-free metrics (Lowe et al., 2017; Tao et al., 2017) would be interesting, the focus of the current work is on improving reference-based evaluation via unsupervised reference augmentation. While reference-less metrics offer convenience to work with zero or a very small number of references,

Type	Text
CONTEXT	A: How may I help you ?
SINGLE	B: I'm having a problem.
MULTI	B: You can help me with this problem. ‖ B: You can tell me how to get to customer service.
RETRIEVAL	B: I have a problem. ‖ B: There is a problem with my check.
COMMONSENSE	B: I want to find information. ‖ B: I want to ask question. ‖ B: I want to make appointment.
CONTEXT	B: It can be solved by drawing a draft on us at 90 days sight. A: What about a draft at 120 days sight ?
SINGLE	B: All right. But we demand the draft be accepted by a bank acceptable to us.
MULTI	B: We'd like the matter resolved sooner. ‖ B: We can do that, but there will be a higher interest rate.
RETRIEVAL	B: Well, that's a lot of time to wait for the draft to be drawn.
COMMONSENSE	B: I want to sign the contract. ‖ B: I will look at the draft sheet.

Table 3: Example showing the automated responses returned by different sub-components of SCARCE. Multiple responses from the same sub-component are separated by '‖'.

reference-based metrics can be advantageous on several fronts. Reference-based evaluation can be more interpretable under certain situations by identifying the reference which matches the most with a given system output. Reference-based evaluations allow for easy incorporation of additional references – in contrast, many learned model-based metrics will require retraining if additional annotations become available.

5 Related Work

Prior work explores many ways to improve over single-reference evaluation without collecting multiple ones. Fomicheva et al. (2020) obviate need for multiple references in MT by generating many "alt-hypotheses" via test-time dropout from the same model. Sai et al. (2020a) and Gupta et al. (2019) collect additional manually annotated responses for dialog contexts. Compare to them, our method of automatically collecting additional references automatically is more scalable.

Automatic data augmentation in NLP has largely been used for increasing training data (Feng et al., 2020; Wei and Zou, 2019; Feng et al., 2021). In this work, we use retrieved dialog instances and commonsense knowledge base to augment reference set for a given dialog context. Δ-Bleu (Galley et al., 2015) and uBLEU (Yuma et al., 2020) also use retrieval to produce pseudo-references for dialog response evaluation. Compared to Δ-Bleu and uBLEU, our work is different since we utilize commonsense knowledge base and perform contextual adaptation. Prior work in dialog response generation has explored the use of commonsense knowledge base (Majumder et al., 2020) as well as retrieval (Song et al., 2016; Majumder et al., 2021) – in contrast, our focus is on augmenting reference set for improving evaluation.

Automatic model-based metrics like ADEM (Lowe et al., 2017) and RUBER (Tao et al., 2017), which incorporate context while scoring for evaluation, at first glance seem to reduce the need for multiple references. However, these metrics have been found to suffer from several peculiar problems. For instance, ADEM can't discriminate between gold responses and certain classes of adversarial negatives e.g reversed gold responses and repeating the context as the response (Sai et al., 2019). Sato et al. (2020) evaluate dialog systems through their ability at selecting valid responses from a semi-automatically curated candidate list. Mehri and Eskenazi (2020b) introduce the unsupervised, reference-free USR metric, which leverages a suite of RoBERTa (Liu et al., 2019) models, each fine-tuned to score one of five dialog aspects e.g *Natural* and *Uses Knowledge*. Mehri and Eskenazi (2020a) further expand their USR metric to eighteen aspects from the initial five.

6 Conclusion

In this work, we demonstrate how existing knowledge sources can be used to construct a diverse set of references in an automated and scalable manner. The resulting reference set demonstrates high correlation with human ratings of system outputs.

In future, we plan to incorporate other commonsense types into SCARCE, such as social (Sap et al., 2019b) and moral (Forbes et al., 2020). We also hope to explore human-in-the-loop setups which build on SCARCE to collect even better references.

Acknowledgements

We thank anonymous ACL reviewers for insightful comments and feedback. We thank Prakhar Gupta (Gupta et al., 2019) for useful discussions.

Ethics Statement

Our preference ratings are collected over source content from an already existing, publicly available and widely used dataset i.e DailyDialog (Li et al., 2017) We neither solicit, record or request any kind of personal or identity information while collecting our ratings. Our work primarily performs experiments on dialog in English (Bender and Friedman, 2018). Dialog models are known to suffer from biases learnable from dialog training data, such as gender bias (Dinan et al., 2019). However, our work and contribution does not present or release any new models or model checkpoints, and is primarily focussed on making existing evaluation setups better through automated collection of larger reference sets.

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.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: commonsense Transformers for Automatic Knowledge Graph Construction. In *ACL*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. Genaug: Data augmentation for finetuning text generators. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 29–42.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Marina Fomicheva, Lucia Specia, and Francisco Guzmán. 2020. Multi-hypothesis machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1218–1232.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. $\triangle$-bleu: A discriminative metric for generation tasks with intrinsically diverse targets. In *ACL (2)*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P 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.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3294–3302.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

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.

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. *arXiv preprint arXiv:1907.11692*.

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.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294. The Association for Computer Linguistics.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, and Harsh Jhamtani. 2021. Unsupervised enrichment of persona-grounded dialog with background stories. In *ACL*.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*.

Shikib Mehri and Maxine Eskenazi. 2020b. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

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.

Hao Peng, Ankur Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. 2019. Text generation with exemplar-based adaptive decoding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2555–2565.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 794–805. Association for Computational Linguistics.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6220–6227.

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

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2020b. A survey of evaluation metrics used for nlg systems. *arXiv preprint arXiv:2008.12009*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. 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 4453–4463.

Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. 2020. Evaluating dialogue generation systems via response selection. *arXiv preprint arXiv:2004.14302*.

Roger C Schank. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology*, 3(4):552–631.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

I. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Charles Spearman. 1961. The proof and measurement of association between two things.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Hiroaki Sugiyama, Toyomi Meguro, and Ryuichiro Higashinaka. 2019. Automatic evaluation of chat-oriented dialogue systems using large-scale multi-references. In *Advanced Social Interaction with Agents*, pages 15–25. Springer.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. *arXiv preprint arXiv:1701.03079*.

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

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6383–6389.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Tsuta Yuma, Naoki Yoshinaga, and Masashi Toyoda. 2020. ubleu: Uncertainty-aware automatic evaluation method for open-domain dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 199–206.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

A Additional Results

A.1 Additional Correlation Results

Table 4 shows Spearman rank correlation scores with p-values.

A.2 Quality Assessment based on RUBER

As a second, automated way of ascertaining response quality, we use the unreferenced part of the RUBER metric (Tao et al., 2017), which uses a pretrained model to score quality of responses based on context alone. Here, we use the RUBER checkpoint[3] from (Sai et al., 2020a), which first pretrains on a large Reddit dataset, followed by finetuning on DailyDialog. SINGLE and MULTI have a quality of ≈ 0.72, while for RETRIEVAL the values is 0.63 . COMMONSENSE is found to have the most superior quality at 0.82, surpassing even MULTI.

A.3 Diversity of References

We investigate the diversity of the references by computing self-BLEU scores (Zhu et al., 2018) among references from PARAPHRASE vs SCARCE. For fair comparison, we randomly chose 4 references from corresponding method. We observe self-BLEU4 scores of 0.36 for PARAPHRASE compared to only 0.13[4] for SCARCE.

B Additional Details on Context Adaptation

B.1 Templates to convert Knowledge Base Outputs to Full Sentences

Table 5 lists the set of templates and rules used to transform semi-structured COMET outputs to surface natural language forms.

B.2 Unsupervised Decoding Procedure For Context Adaptation

We use the author's own implementation[5] of their DELOREAN decoding algorithm from (Qin et al., 2020). We use default hyperparameters from their implementation, which use the non-finetuned *gpt2-medium* checkpoint as the LM atop which the unsupervised, gradient-based decoding procedure is run. Note that the model parameters are not updated in any way - the gradient computation and updates here are happening w.r.t the states, or more specifically, the state activation. More specifically, authors propose an iterature procedure wherein they alternatively perform forward and backward passes. In the forward pass, the current output Y is updated as per the likelihood of the underlying decoder. In the backward pass, the output is updated to be as similar as possible to the sentence Z from the knowledge source using back-propagation. However, since Y is discrete, we maintain a soft representation $\widetilde{Y}$ of the output Y wherein $\widetilde{Y}_i$ represents the logits at the i^{th} position as per the underlying decoder. Next, we shall describe the backward and forward passes of the iterative procedure:

1: In backward pass, we update logits based on the gradients of a content-matching loss function $\nabla_{\widetilde{Y}} L(\widetilde{Y}_{t-1}, Z)$ giving backward logits $\widetilde{y}_t^b$

2: Next, we perform forward pass using the underlying decoder for steps 1 to N. During forward pass at step t, we compute the logits $\widetilde{y}_t^f$ based on left context i.e. X and $Y_{<t}$. Next we perform weighted averaging of the forward and backward logits at step t to arrive at the final logits to be used for the next time step in forward pass.

$\widetilde{Y}_i$ is initialized by performing a forward pass conditioned only on X as per greedy decoding. We alternatively perform backward and forward passes till convergence. Final response is obtained via the resulting logit outputs $\widetilde{Y}$.

Specifically, we use their "counterfactual" setup, where an ending e_{old} is adapted from its old context c_{old} to an altered, new context c_{new}, generating a new, predicted ending $\hat{e}_{new}$. In our case, c_{new} is the dialog context for the turn under evaluation d_t^{past}. In the RETRIEVAL case, c_{old} is the context of the retrieved candidate turn $x_{t'}^{past}$. For the COMMONSENSE case, c_{old} is also our current context, i.e the same as c_{new} - we're simply attuning the already drawn inference better to the current context.

C Retrieval Similarity Function - Details

Consider a dialog d , broken up by turns as $\{C_1 \dots C_t, C_{t+1}{=}d_t^{resp}, C_{t+2} \dots C_T\}$, where $t{+}1$ denotes the turn currently under evaluation. For the context-response $C_t^1, \hat{r}_t$ pair to be evaluated, we retrieve pseudo-references based on a combination of of a) Past $d_t^{past} = C_t^{t-L_b}$ b) Gold response d_t^{resp} c) Future $d_t^{future} = C_{t+2+L_f}^{t+2}$. L_b and L_f are past and future context windows. Our retrieval similarity function is a sum of the log scores between each corresponding element of the turn under evaluation with the candidate turn.

[3] tinyurl.com/ynqd54tt

[4] Note that lower self-BLEU denotes more diverse

[5] tinyurl.com/2lqp9z6s

Spearman Rank Correlation (p-values)

Setup	1 human written reference			4 human written references		
Dataset	SINGLE (Li et al., 2017)	PARAPHRASE -SINGLE	SCARCE -SINGLE(Ours)	MULTI (Gupta et al., 2019)	PARAPHRASE -MULTI	SCARCE -MULTI(Ours)
BLEU-4	0.093 (0.04)	0.135 (0.00)	0.302 (0.00)	0.281 (0.00)	0.269 (0.00)	0.357 (0.00)
BLEU-3	0.055 (0.22)	0.105 (0.02)	0.291 (0.00)	0.243 (0.00)	0.238 (0.00)	0.345 (0.00)
BLEU-2	0.040 (0.37)	0.082 (0.07)	0.275 (0.00)	0.203 (0.00)	0.206 (0.00)	0.327 (0.00)
BLEU-1	0.024 (0.59)	0.062 (0.17)	0.250 (0.00)	0.191 (0.00)	0.178 (0.00)	0.295 (0.00)
ROUGE-L	0.071 (0.11)	0.088 (0.05)	0.259 (0.00)	0.197 (0.00)	0.196 (0.00)	0.317 (0.00)
METEOR	0.106 (0.02)	0.094 (0.04)	0.243 (0.00)	0.227 (0.00)	0.217 (0.00)	0.299 (0.00)
EmbeddingAvg	0.030 (0.50)	0.015 (0.73)	0.025 (0.58)	0.099 (0.03)	0.096 (0.03)	0.079 (0.08)
SkipThought	-0.003 (0.95)	-0.033 (0.46)	0.087 (0.05)	0.065 (0.15)	0.053 (0.24)	0.129 (0.00)
BERT-Prec	0.270 (0.00)	0.279 (0.00)	0.378 (0.00)	0.319 (0.00)	0.322 (0.00)	0.407 (0.00)
BERT-Rec	0.096 (0.03)	0.094 (0.04)	0.240 (0.00)	0.232 (0.00)	0.212 (0.00)	0.304 (0.00)
Max. value	0.270	0.279	0.378	0.319	0.322	0.407

Table 4: Utterance level Spearman Rank Correlation (Spearman, 1961) with p-values. (1) SCARCE-SINGLE augments the original single human written response (SINGLE) in DailyDialog dataset (Li et al., 2017) using proposed method. It leads to large improvements in correlations across most of the metrics, when compared to SINGLE. (2) SCARCE-MULTI augments the MULTI dataset, again leading to improvements in correlations to human ratings, especially for BLEU and BERT-Prec metrics. Additionally, we note that almost all of the correlation values with SCARCE-MULTI are statistically significant with $p < 0.05$.

Condition	Action
Type is OEFFECT	Prepend 'I feel'
Example:	
OEFFECT (excited)	=> 'I feel excited.'
Type is OWANT	Prepend 'I'
Example:	
OWANT (to thank personx)	=> 'I want to thank personx.'
Type is OREACT	Prepend 'I will'
Example:	
OREACT (have a party)	=> 'I will have a party.'
Word PERSONX	Replace with 'you'
Example:	
i thank PERSONX.	=> 'I thank *you*.'

Table 5: Templates and rules to transform semi-structured COMET outputs to surface NL forms.

$$Sim(d_t, x_{t'}) = \log S_{bm25}(d_t^{past}, x_{t'}^{past}) + \log S_{bm25}(d_t^{resp}, x_{t'}^{resp}) + \log S_{bm25}(d_t^{future}, x_{t'}^{future})$$

We set $L_b = L_f = 2$ without specific tuning, as an intuitive tradeoff between enough specificity and enough possibility of relevant candidates.

BM25 (Robertson et al., 1995) or "Best Match 25" is a tfidf like similarity. Its specific form is:

$$S_{BM25}(q, d) = \sum_{w_i \in q} \log(\frac{N}{df_i}) \frac{(k_1 + 1) tf_i}{k_1((1 - b) + b\frac{dl}{avdl}) + tf_i}$$

Here, tf_i and df_i are the term frequency in the current document and the document frequency (in the corpus). N is corpus size, while dl and $avdl$ are current and average document lengths. b controls extent of document length normalization, while k_1 controls effect of term frequency. With $b = 0$ and $k_1 \to \infty$, this reduces to simple tfidf . Here, we use default gensim values, $b = 0.7$, $k_1 = 0.5$

D Qualitative Examples

In Tables 6, we list some examples, each illustrating a turn of a test dialog with its immediate past, future, the four additional human references from (Gupta et al., 2019) (shown under MULTI 2,3 and MULTI 4,5), followed by automated response sets from different sub-components of SCARCE.

D.1 Before/After CONTEXTADAPT

In Example *4-4* of Table 6, we can observe how *"Yes , I'm young , and unmarried . It's no problem for me to travel frequently ."* gets context-adapted (shown as +CA, short for CONTEXTADAPT) to *"Yes , I'm able to understand English. It 's not that I don't understand English ."* which indeed does match the preceding dialog better. Similarly, in Example *50-2* of Table 6, we can see how *"Well, that might be acceptable if you handle insurance fees"* is modulated stronger to the context which asks about duration, getting adapted to *"Well, that's a lot of time to wait for the draft to be drawn."*. Note that we omit this row for the examples where it simply leaves the input unchanged, or produces outputs which are noticeably unfaithful or ill-formed.

D.2 Complementarity of Components

Sometimes, a component may suffer from example specific issues e.g In Example *35-2*, the COMMONSENSE approaches misinterpret what is a driving

ticket in the context of the dialog as an event ticket, drawing inferences accordingly. However, even in such cases, the other component salvages the situation and ensures overall response set remains healthy - e.g, here, RETRIEVAL produces pertinent responses like *Could you tell me how you dealt with it this time?*, *No I haven't. What about you ?* etc. In Example *10-3*, we see the opposite situation, where the responses from RETRIEVAL are somewhat less relevant, but COMMONSENSE produces very pertinent responses such as *"i decline the date"* and *"I go on another date"*

E Human Evaluation Details

E.1 Quality of References

The quality of references were judged by two graduate students from a university where the medium of instruction is English. The annotators were requested to ignore minor grammar issues, and focus more on the content of the response.

F Computing Details

The GPUs used for COMMONSENSE and CONTEXTADAPT experiments were a Geforce Rtx 2080 and TitanX Pascal respectively.

No	Type	Text
0-5	CONTEXT	A: I also have blow if you prefer to do a few lines.
	FUTURE	A: come on man ! i even got dope and acid ! try some !.
	SINGLE	B: No, I am ok, really.
	MULTI 2,3	B: what is your price ? ; B: no thanks, I'm out of cash
	MULTI 4,5	B: yeah , what have you go got ?; B: maybe tomorrow , i 'm running late .
	RETRIEVAL	B: Well , I guess Zina figured you couldn't blow it if you were in the hotel room .
	RETRIEVAL W CA	B: I have a few lines of blow , but it is not necessary in the way of the rest of the script .
	COMMONSENSE	B: i want to go to bathroom; B: i want to play game; B: i want to have rest; B: i want to sleep
1-1	CONTEXT	B: The taxi drivers are on strike again A: What for?
	FUTURE	A: it is really a hot potato .
	SINGLE	B: They want the government to reduce the price of gasoline .
	MULTI 2,3	B: they want more pay for their time .; B: they are not happy with their permit related expenses .
	MULTI 4,5	B: they want healthcare and benefits .; B: i don't know . i just know there are no cabs in the city
	RETRIEVAL	B: Well . Don't worry about it , I will be happy to share the taxi with you .
		B: I think the price is a little high , can't you reduce it?
	RETRIEVAL W CA	B: Well , they are not going to be able to get their fares back .
		B: I think the reason is a little more complicated than that .
4-4	CONTEXT	A: Do you think you can make yourself easily understood in English ?
	SINGLE	B: yes , in most circumstances .
	MULTI 2,3	B: my english is fine . my keyboard sometimes just acts up . ; B: yes, and i'm constantly trying to improve my communication.
	MULTI 4,5	B: i do n't know . i may be to nervous and then i 'll mess up . ; B: yes , i believe so . i 've been speaking english since birth .
	RETRIEVAL	B: Yes , I'm young , and unmarried . It's no problem for me to travel frequently .
	RETRIEVAL W CA	B: Yes , I'm able to understand English . It 's not that I don't understand English .
10-3	CONTEXT	B: Hold on , please . Let me check it for you . Yes , you're right . You will keep it for 3 days .
		A: Well , now I want to change the date from 24th to 28th .
	SINGLE	B: ok , that shall be arranged .
	MULTI 2,3	B: we can easily do that for you . ; B: i 'm sorry but someone's reserved the room for 28th . i can offer you a different room ?
	MULTI 4,5	B: i can extend your stay time but it may cost more since it is going into the holiday . ;
		B: i can give you a different double room for the 28th at a discounted price .
	RETRIEVAL	B: All right . May I have your name ? ; B: Apparently there is only $ 57 left in your account . ;
		B: Here you are . What is the rate ?; B: I prefer not to move first .
	COMMONSENSE	B: i decline the date ; B: i go on another date; B: i go on a date;
		B: i get nervous; B: i get rejected; B: i get dumped; B: i feel sad
23-10	CONTEXT	B: I mainly eat baked chicken , because there's not a lot of fat . A: that does sound pretty good .
	SINGLE	B: i know it does , and that 's because it is .
	MULTI 2,3	B: yes , it is very good and pretty healthy .; B: it is good , you should come over and i will make you dinner .
	MULTI 4,5	B: thanks , it 's been a struggle , but i am starting to see results . ;
		B: i appreciate your kind words , i am getting used to eating more healthy .
	RETRIEVAL	B: You should try it . You won't be disappointed .; B: Do you have any plans for tomorrow ? ; B: That's it ? ;
		B: Just about . ; B: I like fish . Please bring me an order of that .
	COMMONSENSE	B: i want to cook dinner; B: i want to go to movie; B: i want to write letter ;
		B: i will open your mouth; B: i will think about what you want; B: i will drink some water;
35-2	CONTEXT	B: I've never gotten one . Have you ? A: I got a ticket yesterday .
	FUTURE	A: i parked in the red zone .
	SINGLE	B: really ? what did you get one for ?
	MULTI 2,3	B: what was it for ? ; B: how much is it ?
	MULTI 4,5	B: do you resent you got a ticket ? ; B: are you going to appeal the ticket ?
	RETRIEVAL	B: I've gotten a few .; B: No , I haven't . What about you ? ;
		B: Could you tell me how you dealt with it this time ? ;
		B: I will explain it in detail in the next class , could you preview it ?
	COMMONSENSE	B: i give ticket to you; B: i give ticket to the cashier; B: i give ticket to the clerk.;
		B: i feel happy .; B: I feel excited .; B: i feel annoyed . ; B: I feel disappointed .;
		B: i see what else they can do .; B: i see what else they can do to get the ticket; i go to the event

Table 6: Example context-response pairs from the test split of DailyDialog, showing the automated responses returned by different sub-components of SCARCE. CONTEXTADAPT is shortened to CA for brevity.