# QualityAdapt: an Automatic Dialogue Quality Estimation Framework

**John Mendonça**[1,2,*], **Alon Lavie**[3] and **Isabel Trancoso**[1,2]
[1]INESC-ID, Lisbon
[2]Instituto Superior Técnico, University of Lisbon
[3]Unbabel, Pittsburgh
{john.mendonca, isabel.trancoso}@tecnico.ulisboa.pt
alon.lavie@unbabel.com

## Abstract

Despite considerable advances in open-domain neural dialogue systems, their evaluation remains a bottleneck. Several automated metrics have been proposed to evaluate these systems, however, they mostly focus on a single notion of quality, or, when they do combine several sub-metrics, they are computationally expensive. This paper attempts to solve the latter: **QualityAdapt** leverages the Adapter framework for the task of Dialogue Quality Estimation. Using well defined semi-supervised tasks, we train Adapters for different subqualities and score generated responses with AdapterFusion. This compositionality provides an easy to adapt metric to the task at hand that incorporates multiple subqualities. It also reduces computational costs as individual predictions of all subqualities are obtained in a single forward pass. This approach achieves comparable results to state-of-the-art metrics on several datasets, whilst keeping the previously mentioned advantages.

## 1 Introduction

Open-domain neural dialogue systems have increasingly drawn attention in Natural Language Generation (NLG). These systems, colloquially known as Chatbots, take advantage of large-scale training of complex models, making them increasingly more humanlike (Zhang et al., 2020; Adiwardana et al., 2020a; Roller et al., 2021). A crucial step in the development of a dialogue system is its evaluation. The community has identified multiple characteristics of what constitutes a high-quality dialogue. These include comprehensible, fluent, empathetic, relevant and interesting, among others. The precise definition is often challenging to define and is application dependent.

The current trend is to train models to evaluate responses under various aspects. These learning-based metrics either (1) map overall quality to a single defined aspect such as Sensibleness (is the response adequate given the context) or (2) leverage several individual models to cover a wider range of quality aspects (subqualities). Both have their drawbacks: in the first approach, the use of a single notion of quality limits the overall understanding of model performance and consequently its applicability to other domains; in the second approach, the need to individually train several models is both time and resource consuming, possibly duplicating model parameters that could be shared, such as feature representations.

This paper proposes QualityAdapt[1], an automatic dialogue quality estimation framework that leverages the Adapter paradigm (Houlsby et al., 2019a) to train individual Adapters on different dialogue subqualities. Then, AdapterFusion (Pfeiffer et al., 2021) combines the knowledge of the individual Adapters for the downstream task of overall quality estimation. This allows for a system that is both extensible (by including different subqualities) and less resource-intensive (by sharing most of the pretrained model parameters). Experimental results show that QualityAdapt achieves comparable correlations with human judgements when compared to other state-of-the-art metrics.

## 2 Background

### 2.1 Automatic Quality Estimation Metrics

Word-overlap metrics, such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), are a popular choice to evaluate dialogues as they are used to evaluate machine translation and summarization models and are easy to employ. These metrics assume valid responses have significant word-overlap with the ground truth. However, this is not a valid assumption: there are many equally good responses for a single utterance. As

* Corresponding author

[1]Model parameters and codebase are available at: github.com/johndmendonca/qualityadapt.



*Proceedings of the SIGdial 2022 Conference*, pages 83–90
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 

such, the correlation with human judgements is very low for these metrics (Liu et al., 2016), and they cannot be used to evaluate models in an online setting, where a gold-response is not available.

Earlier learned metrics such as ADEM (Lowe et al., 2017) and RUBER (Tao et al., 2018) explicitly predict human annotations by initialising pretrained RNN response generators. In both cases, a reference response is used to score the candidate response. As such, these metrics still suffer the same issues as word-overlap metrics.

More recently, open-domain automatic dialogue quality estimation has concentrated on reference-free methods. Most metrics focus on evaluating a single notion of quality such as Engagement (Ghazarian et al., 2020), Sensibleness (Dziri et al., 2019; Huang et al., 2020) or Human-likeness (Gao et al., 2020). Metrics such as USR (Mehri and Eskenazi, 2020b), USL-H (Phy et al., 2020) and Deep AM-FM (Zhang et al., 2021b) combine predictions of individual sub-metrics obtained from Language Models.

### 2.2 Adapters

Adapters in NLP (Houlsby et al., 2019b) have been introduced as an alternative to the full model fine-tuning strategy. They consist of a small set of additional trainable parameters added between layers of a pretrained network. These consist of feed-forward layers with normalizations, residual connections, and projection layers. The weights are trained during fine-tuning for a given task, while the pretrained parameters of the large model are kept frozen. This strategy allows for parameter sharing by training different task and language specific Adapters using the same model. Furthermore, previous work has shown that Adapters achieve comparable performance to full fine-tuning (Pfeiffer et al., 2020a, 2021), despite the primary focus being geared towards parameter efficiency.

**AdapterFusion (Pfeiffer et al., 2021)** proposes improving downstream task results by transferring task specific knowledge obtained from training Adapters on supporting tasks. The architecture takes inspiration from the attention mechanism (Vaswani et al., 2017), and consists of learnable weights Query, Key, and Value: the Query consists of the pretrained transformer weights; the Key and Value take as input the output of the respective Adapters. The dot product of the query with all the keys is passed into a softmax function, which learns to weight the Adapters with respect to the context. Therefore, the goal is to learn a parameterized mixer of the available trained Adapters.

## 3 QualityAdapt

QualityAdapt trains individual Adapters for each subquality and composes them using AdapterFusion for the task of overall quality estimation. In both the subquality and overall quality tasks, it returns a score that is obtained by combining a transformer encoder with a regression head on top. During inference, individual subquality predictions can be obtained in a single forward pass by parallelising their respective heads.

**Encoder** In our experiments, RoBERTa-large (Liu et al., 2019) is used to encode the context-response pair. In the tokenization step, we add for each utterance a token representative of the speaker. This added information lets the network identify the response's speaker, which in turn allows it to pay more attention to utterances from this speaker in the context if needed.

**Compositionality** Training AdapterFusion for the downstream task of overall quality estimation is a supervised task. As such, quality annotated data in terms of overall quality is required. However, the amount of annotations required for the Fusion training step is much smaller when compared to fully fine-tuning a Language Model with this data. As a proof of concept, we composed two Adapters in this paper: **U-Adapter**, for Understandability, and **S-Adapter** for Sensibleness.

**U-Adapter** An understandable response is one that can be understood without context. Such responses may contain minor typos that do not hinder the comprehension of the response. Mehri and Eskenazi (2020b) evaluates this sub-metric by calculating the likelihood of the response using a Masked Language Modelling (MLM) metric. In this paper, we follow the approach used by Phy et al. (2020) and initially proposed by Sinha et al. (2020). A model is trained to differentiate between positive samples and synthetic negative samples. Positive samples are perturbed by randomly applying one of the following: (i) no perturbation, (ii) punctuation removal, (iii) stop-word removal. Negative samples are generated by randomly applying one of the following rules: (i) word reorder (shuffling the ordering of the words); (ii) word-drop; and (iii) word-repeat (randomly repeating words).

**S-Adapter** A sensible response is one that takes into account its preceding context. The task of predicting sensibleness can be considered a binary Next Sentence Prediction (NSP) task, distinguishing a positive example (the subsequent utterance) from a semantically negative one (a random utterance from a response pool obtained from the dataset). Many dialogue quality estimation metrics leverage the NSP task when training their models for quality estimation (Zhao et al., 2020; Zhang et al., 2021a; Phy et al., 2020; Mehri and Eskenazi, 2020b).

## 4 Experiments

### 4.1 Datasets

Different data sources are used in the experiments:

**Training** – DailyDialog (Li et al., 2017) is used for the self-supervised training and evaluation of the S and U Adapters. Additionally, the Fusion module is trained using the annotated split by Zhao et al. (2020) (denoted as *DD-Z*).

**Evalution** – The evaluation of the subqualities is done on the data annotated by Phy et al. (2020) (denoted as *DD-P*). QualityAdapt's extensibility is also evaluated on different overall quality annotated datasets:

- TopicalChat (Gopalakrishnan et al., 2019) and PersonaChat (Zhang et al., 2018), which were annotated by Mehri and Eskenazi (2020b) and denoted in this work as *USR-TC* and *USR-PC*, respectively;
- *DSTC6* (Hori and Hori, 2017);
- *FED* (Mehri and Eskenazi, 2020a).

A more detailed overview of these datasets can be found in Appendix A.

### 4.2 Baselines

**USR (Mehri and Eskenazi, 2020b)** leverages several Language Models to measure dialogue properties. These include: *Fluency*, measured using masked language modelling (MLM) objectives; *Relevance*, using a dialog retrieval model and *Uses Knowledge*, measured using a fact-to-response selection model. Overall quality prediction is obtained using a Linear Regression model.

**RoBERTa-eval (Zhao et al., 2020)** proposes an evaluator that produces an encoding vector given a context and a response, and then calculates its score via an MLP with a sigmoid function. The model takes the pretrained transformer and primes it on an NSP task with in-domain data using Negative Sampling, which offsets the lack of annotated data. A final finetuning is done for quality prediction.

**USL-H (Phy et al., 2020)** combines three models trained with different objectives: Valid Utterance Prediction (BERT-VUP), Next Sentence Prediction (BERT-NSP), and BERT-MLM. The BERT-VUP model determines whether a response is valid and grammatically correct. The BERT-NSP model and BERT-MLM models are trained with self-supervised objectives to evaluate the sensibleness and the likelihood of a given response.

### 4.3 Subquality Estimation

| | | Pearson | Spearman |
|---|---|---|---|
| Understand. | BERT-MLM | -0.16 | *0.01* |
| | BERT-VUP | 0.26 | *0.14* |
| | USR-MLM | *0.01* | *0.11* |
| | RoBERTa-large | **0.35** | *0.18* |
| | U-Adapter | 0.32 | **0.21** |
| Sensible | BERT-NSP | 0.63 | 0.61 |
| | USR-DR (x=c) | 0.54 | 0.47 |
| | RoBERTa-large | 0.61 | 0.65 |
| | S-Adapter | **0.68** | **0.67** |

Table 1: Correlation for Understandability and Sensibleness subquality between human annotations and automatic metrics. Best results are denoted in **bold**, *italic* identifies $p > 0.01$.

The test set results on the DailyDialog dataset for the Understandability and Sensibleness subqualities are presented in Table 1. Here, we evaluate the correlation between the average human annotation and the model prediction. For fair comparison, we also include the results with a fully finetuned RoBERTa-large model. With respect to the estimation of Understandability, U-Adapter outperforms the models proposed by USR (USR-MLM perplexity) and USL-H (BERT-VUP). Similar results are observed on the Sensibleness task, where both RoBERTa and S-Adapter outperform both USL-H (BERT-NSP) and USR baselines. These results confirm Adapters are a valid substitute to fully finetuned models for the task of subquality estimation.

### 4.4 Overall Quality Estimation

In the overall quality prediction task, we compare the different metrics on all datasets. Results in Table 2 show that, on average, the S+U metric outperforms all other metrics on these datasets. As expected, all models obtain the best performance

| | DD-Z | | DD-P | | USR-TC | | USR-PC | | DSTC6 | | FED | | Avg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pr. | Spr. | Pr. | Spr. | Pr. | Spr. | Pr. | Spr. | Pr. | Spr. | Pr. | Spr. | Pr. | Spr. |
| USR | 0.38 | 0.39 | 0.51 | 0.48 | **0.41** | **0.42** | 0.44 | 0.42 | 0.18 | 0.17 | 0.11 | 0.12 | 0.34 | 0.33 |
| USL-H | *0.25* | *0.26* | 0.63 | 0.64 | 0.32 | 0.34 | **0.50** | **0.52** | 0.22 | 0.18 | 0.20 | 0.19 | 0.35 | 0.36 |
| RoB-eval | 0.64 | 0.66 | 0.73 | 0.74 | 0.22 | 0.22 | 0.34 | 0.33 | 0.28 | 0.29 | **0.29** | **0.26** | 0.42 | 0.41 |
| S+U | **0.73** | **0.74** | 0.76 | **0.76** | 0.29 | 0.29 | 0.36 | 0.36 | **0.43** | **0.42** | 0.27 | 0.23 | **0.47** | **0.47** |
| -U Adapter | 0.67 | 0.69 | **0.80** | **0.76** | 0.28 | 0.30 | 0.37 | 0.37 | 0.39 | 0.40 | *0.17* | *0.13* | 0.45 | 0.44 |
| -Speaker | 0.62 | 0.65 | 0.67 | 0.70 | 0.33 | 0.33 | 0.36 | 0.36 | 0.33 | 0.31 | 0.20 | 0.20 | 0.42 | 0.42 |
| -Fusion | 0.60 | 0.54 | 0.72 | 0.73 | 0.20 | 0.23 | 0.37 | 0.34 | 0.36 | 0.33 | 0.17 | 0.21 | 0.40 | 0.40 |
| S+U+E | 0.68 | 0.70 | 0.76 | 0.73 | 0.18 | 0.19 | 0.36 | 0.36 | 0.36 | 0.36 | 0.18 | 0.14 | 0.42 | 0.41 |

Table 2: Correlation for Overall Quality between human annotations and automatic metrics. Best results are denoted in **bold**, *italic* identifies $p > 0.01$. Baseline results are obtained using codebase provided by Yeh et al. (2021).

when evaluated on both DD test sets. Lowest results are obtained on the FED dataset, which contains responses from advanced chatbots, and are therefore more difficult to identify as being low-quality. This underlines the importance of including more subqualities for dialogue evaluation, as contemporary chatbots achieve human performance on typical subqualities such as sensibleness and understandability. This in turn makes them insufficient to discriminate between good and bad responses. However, finer-grained submetrics do not have an obvious mapping to semi-supervised data collection methods, and are therefore discarded due to the lack of sufficient annotated data to fully train models.

### 4.5 Ablation Studies

**Single Adapter Finetuning** In this experiment, we verify the effectiveness of having several Adapters trained on different objectives contributing to the performance of the downstream task. To evaluate this, the U-Adapter and the Fusion module is discarded and the S-Adapter is further finetuned with the quality annotated data (denoted in Table 2 as -U Adapter). On average, dropping the U-Adapter reduces relative performance by 5%.

**Removing Speaker Tokens** We compare the performance of S+U without the speaker tokenization (denoted in Table 2 as -Speaker). Results show the removal of these tokens reduces performance on all datasets except on USR-PC and USR-TC. This may indicate the topic shift between speakers is small and as such "who said what" is inconsequential to sensibleness.

**Removing Adapter Fusion** The contribution of AdapterFusion for the task of quality estimation is assessed by comparing S+U against a Linear Regression model that receives as input the predictions of the individual qualities obtained by the trained Adapters (denoted in Table 2 as -Fusion). The regression model is trained using the same annotated data split as AdapterFusion. Overall, the regression model yields worse results when compared against AdapterFusion. This underlines the power of composition using Fusion, leveraging the learned parameters of the trained Adapters instead of just their prediction.

### 4.6 Emotion Adapter

We posit the emotion conveyed by the agent during the conversation should positively correlate with overall quality annotations: responses that display happiness and excitement are expected to have a positive impact in the dialogue and therefore should favour higher quality annotations when compared to responses that portray neutral, or negative emotions. This was the basis for adding an Emotion Adapter to S+U, denoted S+U+E. The Adapter was trained on the DailyDialog corpus, using the same training parameters as the S and U Adapters, and a Weighted Cross Entropy Loss. A Macro-F1 of 45.00 is achieved on the test set. The inclusion of the emotion Adapter fails to outperform S+U. Our initial hypothesis is that this is due to generative models being conditioned to respond with positive emotions. We leave further investigation of these results for future work.

## 5 Prediction Compute

One of the motivations of the QualityAdapt framework is it's computational efficiency. We present average sample predictions per second on the test set using a single RTX 3070Ti 8BG GPU, together with size of the **metric's unique parameters** on Table 3. For the baseline methods, the transformer model is fully fine-tuned and therefore the full model in included; for the Adapters, only the Adapter, the fusion layer and corresponding heads are included in the calculation. We note that a full transformer model (RoBERTa-base/large) is

| Metric | Samples/s | Model Params |
|---|---|---|
| USR | 22.44 | 4.2 GB |
| USL-H | 10.83 | 3.9 GB |
| RoBERTa-eval | 79.11 | 3.2 GB |
| S (large) | 59.67 | 17.1 MB |
| S+U (base) | 107.29 | 168.8 MB |
| S+U (large) | 59.11 | 319.1 MB |
| S+U+E (large) | 59.24 | 332.1 MB |

Table 3: Prediction loop compute on DD-Z (250 samples). For the QualityAdapt models, (base/large) denote the transformer model's size.

still required for inference in QualityAdapt. However, the sharing of its weights is simplified.

As expected, the forward pass on several transformer models decreases runtime performance when compared to a single forward pass, even when using larger models (USR and USL-H metrics are based on the RoBERTa and BERT-base models, respectively). When comparing between the different larger models, we can see that the inclusion of the Adapter model decreases run-time performance by 25%. However, both the fusion module and the inclusion of more Adapters does not significantly affect performance.

## 6 Conclusions

This paper presents QualityAdapt, a framework for automatic dialogue quality estimation. We show the composition of Sensibleness and Understandability Adapters for the downstream task of quality estimation outperforms, on average, the performance of robust baselines, including those that take advantage of subquality composition. However, QualityAdapt only requires a single forward pass on a Language Model to produce predictions for overall quality, thus reducing computational complexity.[2]

Current research in dialogue focuses mostly on monolingual chatbots, typically in English. Multilingual LMs such as XLM-RoBERTa (Conneau et al., 2020) can be used to extract utterance representations directly in the target language after fine-tuning. However, this approach would still be somewhat limited by the lack of multilingual annotated data. Pfeiffer et al. (2020b) proposes leveraging Adapters for transfer learning in low resource settings by training a stack consisting of the source-language Adapter with a task Adapter. Then, during inference, the source-language Adapter is replaced with the target-language one. We leave these experiments for future work.

[2]The parallel inference of individual Adapters and their fusion using AdapterHub is still WIP.


## Acknowledgements

We would like to thank Patrícia Pereira for helping with emotion experiments, and the reviewers, for their helpful feedback and discussions. This work was supported by national funds through *Fundação para a Ciência e a Tecnologia* (FCT) with references PRT/BD/152198/2021 and UIDB/50021/2020, and by the P2020 program MAIA (LISBOA-01-0247-FEDER-045909).


## References


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

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

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. 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 3806–3812, Minneapolis, Minnesota. 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. In

*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796.

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.

Chiori Hori and Takaaki Hori. 2017. End-to-end conversation modeling track in dstc6. *arXiv:1706.07440*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019a. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning, PMLR*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019b. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

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.

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.

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. 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.

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.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. 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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, 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. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *CoRR*, abs/2106.03706.

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 Fernando D'Haro, Rafael E Banchs, Thomas Friedrichs, and Haizhou Li. 2021b. Deep am-fm: Toolkit for automatic dialogue evaluation. In *Conversational Dialogue Systems for the Next Decade*, pages 53–69. Springer.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

# A Experiments

## A.1 Datasets

**DailyDialog (Li et al., 2017)** is a high-quality human-human open-domain dialogue dataset focused on day-to-day conversations. The dataset consists of 13,118 dialogues and 103,632 utterances. Zhao et al. (2020) (DD-Z) annotates 900 context-response pairs in terms of *Appropriateness* from a pool of responses obtained by negative-sampling response randomly selected from a different dialogue and responses generated by generative models trained on the training split; Phy et al. (2020) (DD-P) collected five responses from two retrieval methods, two generative methods, and one human-generation for 50 contexts. These responses are then annotated in terms of *Understandability*, *Sensibleness*, *Specificity* and *Overall Quality*.

**TopicalChat (Gopalakrishnan et al., 2019)** is a knowledge-grounded human-human conversation dataset that consists of 11,319 dialogues and 248,014 utterances. **PersonaChat (Zhang et al., 2018)** is human-human persona-conditioned conversations that consists of 10,907 dialogues and 162,064 utterances. Mehri and Eskenazi (2020b) (USR-TC) performs human annotation on 60 dialog contexts, with 6 responses per context for TopicalChat (four system outputs, one newly-annotated human output, one original ground-truth response) and five for PersonaChat (USR-PC). Each response was annotated in terms of *Understandability*, *Naturalness*, *Sensibleness*, *Interesting*, *Uses Knowledge* and *Overall Quality*.

**DSTC6 (Hori and Hori, 2017)**, the 6th Dialog System Technology Challenge, used dialog data collected from multiple Twitter accounts of customer service for its conversation modeling track. Each dialogue consisted of real tweets between a customer and an agent. 40,000 responses are obtained from the competing system, all of which are based on the LSTM Seq2Seq model, which are then annotated in terms of overall quality (DSTC-6).

**FED (Mehri and Eskenazi, 2020a)** is constructed by annotating 40 Human-Meena conversations, 44 Human-Mitsuku conversations and 40 Human-Human conversations obtained from Adi-

wardana et al. (2020b). The conversations are annotated with 18 subqualities, at the turn and dialogue levels. In this work we use the turn-level overall quality annotations for evaluation (FED).

### A.2 Training setup and Hyperparamters

This work's codebase uses AdaterHub [3], which is based on HuggingFace Transformers [4]. We train all Adapters using Adam with a learning rate of 1e-4. Training is conducted for 10 epochs, with a batch size of 16, except for the Fusion training, which we set to 8. We experiment different seeds for the Fusion training, and present the best performing one. The best performing model on the evaluation set is selected for testing. Max sequence length was fixed to 128. The regression head consists of 2 layer MLP with a hidden size of 1024. We use the Hyperbolic tangent as the activation function. We use a single Quadro RTX 6000 24GB GPU for training.

[3] https://Adapterhub.ml/

[4] https://github.com/huggingface/transformers