Spurious Correlations in Reference-Free Evaluation of Text Generation

Esin Durmus[1*] **Faisal Ladhak**[2*] **Tatsunori Hashimoto**[1]
[1]Stanford University   [2]Columbia University
esindurmus@cs.stanford.edu faisal@cs.columbia.edu
thashim@stanford.edu

Abstract

Model-based, reference-free evaluation metrics have been proposed as a fast and cost-effective approach to evaluate Natural Language Generation (NLG) systems. Despite promising recent results, we find evidence that reference-free evaluation metrics of summarization and dialog generation may be relying on spurious correlations with measures such as word overlap, perplexity, and length. We further observe that for text summarization, these metrics have high error rates when ranking current state-of-the-art abstractive summarization systems. We demonstrate that these errors can be mitigated by explicitly designing evaluation metrics to avoid spurious features in reference-free evaluation.

1 Introduction

Building reliable automated evaluation metrics is a key factor for quick development of better NLG systems. Recent work has proposed reference-free evaluation metrics as a way to judge the quality of generated outputs without the need for human references (Celikyilmaz et al., 2020). Many of these reference-free evaluations achieve remarkably high correlations with human evaluations, raising hopes that they may soon become a viable alternative to expensive human evaluations (Kryscinski et al., 2020; Goyal and Durrett, 2020; Sinha et al., 2020; Phy et al., 2020; Gao et al., 2020).

However, simply looking at correlation with human scores may not be sufficient to determine the efficacy and robustness of an evaluation metric. In our work, we study recently proposed reference-free evaluation metrics of text summarization and dialog generation. We find that it is possible to achieve similar levels of correlation with human judgment, using simple spurious correlates such as word overlap, length, and perplexity. Furthermore, we find that the learned metrics have a rela-

tively high correlation with the spurious correlates as compared to human scores, which suggests that these metrics may rely heavily on spurious correlations. This may be a potential explanation for the robustness issues that are observed in recent work, despite the seemingly high reported correlations with human judgements (Gabriel et al., 2021; Yeh et al., 2021).

We further analyze reference-free faithfulness evaluation metrics and show that the reliance on spurious correlations leads to errors in model selection and development. First, we show that word overlap, a spurious correlate for the task, does as well as recently proposed reference-free metrics at system-level ranking. Then, we look at rankings amongst systems that are relatively abstractive and faithful, i.e., the current state of the art, and find that these learned metrics perform significantly worse for these systems. This is because word-overlap is not a good measure for ranking these systems in terms of their faithfulness since all of these systems have similarly low word overlap. This suggests that we need metrics that are not overly reliant on word overlap in their faithfulness prediction.

Finally, we explore whether a simple mitigation strategy of adversarially training a faithfulness evaluation metric to avoid spurious correlates can lead to a more robust metric. We find that our adversarially trained metric performs well at overall pairwise ranking while having a significantly lower correlation with the spurious correlate of word-overlap. Crucially, we show that our proposed metric has improved performance in ranking between abstractive and faithful systems, which is a failure mode for existing reference-free faithfulness evaluation metrics.

2 Reference-free Evaluation of Text Generation

We begin by defining the task of reference-free evaluation, as well as the *example-level* and *systems-*

---

*Equal contribution.

level evaluation of these metrics.

We define a reference-free evaluation metric as a function $F(x, y)$ that can assign a quality score to an output sequence y for a given input sequence x. The goal of a reference-free evaluation metric $F(x, y)$ is to assign high scores to desirable outputs y for some attribute, such as the faithfulness of a summary. Measuring the quality of this metric is challenging, and prior work has relied upon correlation to human judgments $H(x, y)$.

Example-level evaluation: A number of existing reference free evaluations rely upon a procedure which we call *example-level* human correlations (Fabbri et al., 2020; Phy et al., 2020; Sinha et al., 2020), which measures the effectiveness of a metric by computing a Pearson or Spearman correlation $\text{corr}_{p_{\text{eval}}}(H(x, y), F(x, y))$ over some sampled evaluation data $p_{\text{eval}}(x, y)$.

System-level evaluation: An alternative approach to evaluation is *systems-level* rankings (Mathur et al., 2020; Kocmi et al., 2021), which we define as the ability to identify which model is better amongst a set of models M. F is evaluated via its accuracy in matching human evaluation H on all pairs $(m_i, m_j) \in M \times M$ where $m_i \neq m_j$.

The definitions of example and system level correlations suggest that evaluations of these metrics may have a strong dependence on the example and systems distributions $p_{\text{eval}}(x, y)$ and M. As an example, consider an evaluation for dialogue response quality. Building a truly accurate predictor for dialogue response quality is challenging, but if $p_{\text{eval}}(x, y)$ consists of all either professionally written examples or ungrammatical nonsense, a simple grammar checker would perform exceedingly well.

This is an instance of what is called a spurious correlate. More formally, we define this as some attribute $S(x, y)$ which is correlated with H in $p_{\text{eval}}(x, y)$ but is not correlated with H for a carefully constructed test distribution $p_{\text{test}}(x, y)$. We say that F is *spuriously correlated* with S if:

1. F and H are highly correlated under $p_{\text{eval}}(x, y)$ but not under $p_{\text{test}}(x, y)$.

2. F remains correlated with S under $p_{\text{test}}(x, y)$.

3   Example-level Analysis of Learned Evaluation Metrics

In this section, we look at example-level Spearman correlations with human judgements for reference-free evaluation metrics that have been proposed for summarization and dialog generation. We compare the metrics to spurious correlates such as word-overlap, length and perplexity, in order to understand whether the metrics can perform better than these simple measures. We also measure to what extent the proposed metrics are correlated with these spurious measures.

3.1   Faithfulness Evaluation in Text Summarization

State-of-the-art text summarization models are capable of producing fluent summaries. However, they suffer from generating information that is not consistent (i.e., unfaithful) with the information in the source article (Cao et al., 2018). Prior work showed that reference-based metrics are not able to capture such consistency errors (Falke et al., 2019). This motivated researchers to build evaluation metrics to capture these faithfulness issues since collecting human evaluations for faithfulness is expensive and time-consuming (Wang et al., 2020; Durmus et al., 2020; Kryscinski et al., 2020; Goyal and Durrett, 2020).

In this section, we analyze recently proposed reference-free faithfulness evaluation metrics and compare their performance against the spurious correlate of word overlap. Furthermore, we analyze the correlation between the learned metrics and word overlap to understand to what extent these metrics rely on spurious correlations. We focus on learned entailment-based faithfulness evaluation metrics due to their high performance in identifying faithfulness issues (Pagnoni et al., 2021). In particular we evaluate FactCC (Kryscinski et al., 2020) and DAE (Goyal and Durrett, 2021), which have been shown to achieve higher example-level correlations with human judgements than existing faithfulness evaluation metrics (Pagnoni et al., 2021).

FactCC. Kryscinski et al. (2020) proposed an entailment-based method where they train a BERT-based model to predict whether or not the source article entails a summary. To train this model, they generate synthetic training data by applying a set of transformations to source article sentences in order to get article, summary pairs. They evaluate their approach on the CNN/DM dataset (See et al., 2017) and report a high accuracy on example-level comparisons on a human-annotated test set.

DAE. Goyal and Durrett (2021) collected human annotations at the word-level and arc-level to study faithfulness at a finer granularity. They also trained
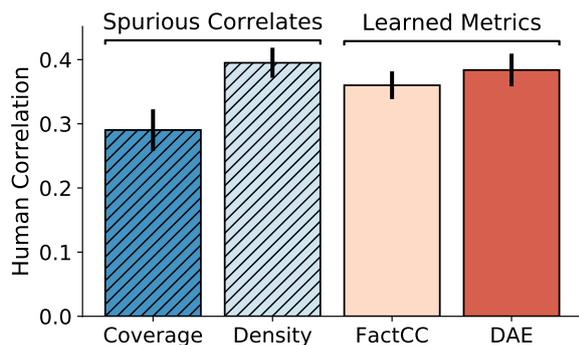
Figure 1: Correlation of the spurious correlates and learned metrics with human scores. Density, a spurious correlate, achieves similar performance as DAE and performs significantly better than FactCC.

Metric	Human	Density
FactCC	0.36	**0.59**
DAE	0.38	**0.76**

Table 1: Correlation of FactCC and DAE scores with humans vs density. Both learned metrics have a significantly higher correlation with density than human scores.

a dependency arc entailment model for faithfulness detection (Goyal and Durrett, 2020). They evaluate on the same test set as Kryscinski et al. (2020) and report improved results over FactCC.

We look at how these learned, reference-free metrics compare with word overlap – a simple spurious correlate. One simple measure of whether a generated summary is faithful is to look at its word overlap with the source article; summaries with a higher word overlap are more likely to be faithful (Ladhak et al., 2021). However, this measure of faithfulness is spurious because it cannot distinguish between faithful and unfaithful summaries that have similar word overlap. In particular, we look at two metrics of word-overlap following Grusky et al. (2018): *coverage* and *density*. *Coverage* measures the percentage of the words in the summary that are also present in the article. *Density* instead looks at the average length of the segments in the summary that are extracted from the article.

Results. We use the large-scale faithfulness human annotations collected by Fabbri et al. (2020) for 16 summarization models on the CNN/DM dataset (See et al., 2017) for our analysis. Figure 1 shows the example-level correlations with human scores for each of the factuality metrics as well as the spurious correlates. We note that *den-*

sity has a similar correlation with human scores as DAE, and is significanlty[1] better than FactCC. This result is alarming because *density* is a spurious correlate, yet it can achieve similar performance as the metrics that have been trained for faithfulness evaluation.

Moreover, we also see that both FactCC and DAE have a significantly higher correlation with *density* than they do with human scores (Table 1). This indicates that these metrics may rely upon spurious correlations and are not yet capturing a deeper understanding of faithfulness.

3.2 Learned Metrics for Dialog Generation

Dialog generation systems need to be able to generate a response given the dialog context. The ability to automatically evaluate the quality of a response is essential for building dialogue systems. Liu et al. (2016) show that referenced-based evaluation metrics do not correlate well with human judgments of response quality. This has led to an increased interest in reference-free evaluation metrics for evaluating dialogue response quality.

Similar to our analysis in § 3.1, we aim to look at recently proposed metrics for reference-free evaluation, along with spurious correlates for dialog response quality, and compare them against human judgments.

DialogRPT. Gao et al. (2020) finetune GPT-2 to predict the different types of human feedback (replies, upvotes, etc.) in Reddit threads and combine these to form a composite score for response quality. They evaluate their approach on the Reddit data that they collected and show that their method achieves higher example-level agreement with human judgments than baseline metrics.

MAUDE. Sinha et al. (2020) propose a model that encodes each utterance in the dialog context using a pre-trained BERT model and leverages the temporal transitions between them to score a response. They add noise to existing dialog responses to create negative examples and train their system to distinguish them from valid responses using noise contrastive estimation (NCE). They evaluate their model on the PersonaChat (Zhang et al., 2018) dataset and report improved example-level Spearman correlation with human judgments compared to existing baseline metrics.

---

[1]All numbers reported in the paper are bootstrap means over 1000 bootstrap samples. We use a one-tailed percentile bootstrap test to determine significance at $\alpha = 0.05$.
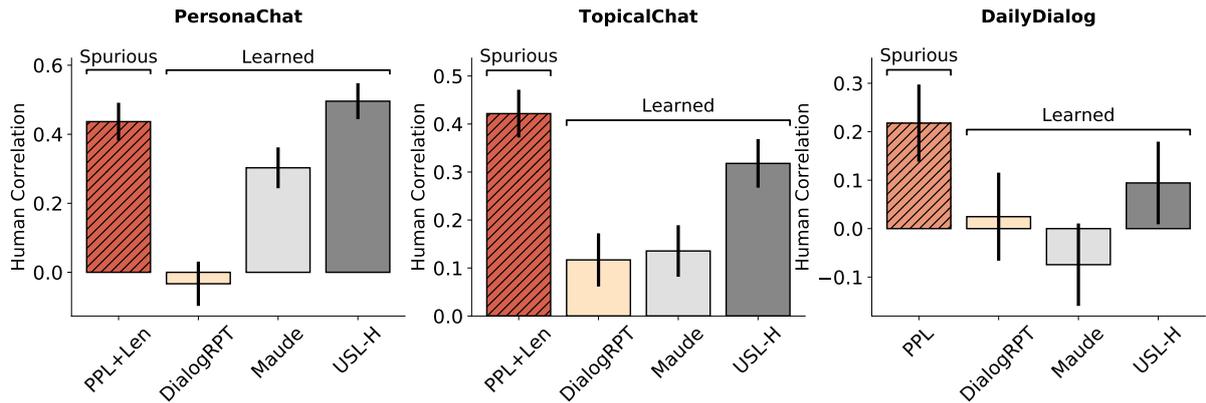
Figure 2: Correlation of the spurious correlates and learned metrics with human scores. PPL+Len represents a simple combination of perplexity (PPL) and length features. The best spurious correlate performs significantly better than all learned metrics on TopicalChat, and performs similarly to the best learned metric on PersonaChat and DailyDialog.

		Human	Perplexity	Length	PPL+Len
	DialogRPT	-0.033	-0.017	**0.086**	0.068
PersonaChat	Maude	0.303	**0.373**	-0.089	0.137
	USL-H	0.496	0.092	**0.506**	0.469
	DialogRPT	0.117	-0.011	0.272	**0.276**
TopicalChat	Maude	0.135	**0.243**	-0.191	-0.148
	USL-H	0.318	0.037	**0.359**	0.355
	DialogRPT	0.025	-0.182	**0.359**	0.270
DailyDialog	Maude	-0.074	-0.076	**0.102**	0.033
	USL-H	0.094	0.048	-0.208	**-0.236**

Table 2: Correlation of the metrics with human scores and spurious correlates. Reference-free evaluation metrics have higher correlation with spurious correlates than the human scores.

USL-H. Phy et al. (2020) decompose response quality into three aspects and train a model to score a response along each of these aspects. They then combine the scores hierarchically into one composite score for response quality. They evaluate their metric on the DailyDialog (Li et al., 2017) dataset and report significantly higher example-level correlations than previous baseline metrics.

MNLI+Adv. Dziri et al. (2021) introduce an entailment-based metric that evaluates the groundedness of a dialog response, i.e., whether the generated response is consistent with the information in the provided external context, such as a Wikipedia article. They trained their metric on automatically generated adversarial data by applying perturbations to the evidence. They further collect human annotations for the various aspects of dialog generation, such as entailment, genericness, etc., and show that their method is more effective in accurately categorizing the generations than existing entailment models.

To assess these metrics, we look at two spurious correlates for dialog quality – perplexity and length of the generated output – as well as a simple combination of two measures. We compute perplexity using a pre-trained GPT-2 language model (Radford et al., 2019). Perplexity (PPL) and length are spurious correlates since they do not account for the dialog context, and therefore it is possible to have high-quality and low-quality responses with similar perplexities/lengths. For groundedness evaluation, we look at the same word overlap measures, as we did for summarization, i.e., *density* and *coverage*, and we measure overlap between the response and the provided external evidence.

Results. We evaluate metrics[2] for response quality estimation on three popular multi-turn dialog datasets – DailyDialog, which contains dialogs

---

[2]We use the code provided by Yeh et al. (2021) for these experiments.

about everyday topics (Li et al., 2017), TopicalChat, which contains dialogs conditioned on a set of 8 broad topics (Gopalakrishnan et al., 2019), and PersonaChat, which contains dialogs conditioned on personas (Zhang et al., 2018).

To evaluate the recently proposed metric for response groundedness, we use human annotations collected by Dziri et al. (2021) on Wizard of Wikipedia (Dinan et al., 2019), a dataset that consists of dialogues conditioned on information from Wikipedia articles. In particular, we use their entailment annotations, where human annotators judge whether or not the external evidence entails a generated response.

Figure 2 shows the correlations with the human scores and the spurious correlates for the dialog generation evaluation metrics. In DialyDialog, we find that perplexity achieves a similar correlation with human judgments as USL-H. In TopicalChat, perplexity or length alone does not beat out any of the learned metrics; however, combining the two measures achieves a significantly better correlation with humans than learned metrics. In PersonaChat, USL-H achieves the highest correlation with human judgment, though the combined PPL+Len score is close. We observe that USL-H is more consistent than the other reference-free metrics and achieves significantly higher correlations with human scores than MAUDE and DialogRPT for PersonaChat and TopicalChat. We further find that the reference-free metrics have a higher correlation with the spurious correlates than the human scores (Table 2), which again suggests that these learned metrics may be relying upon spurious correlations.
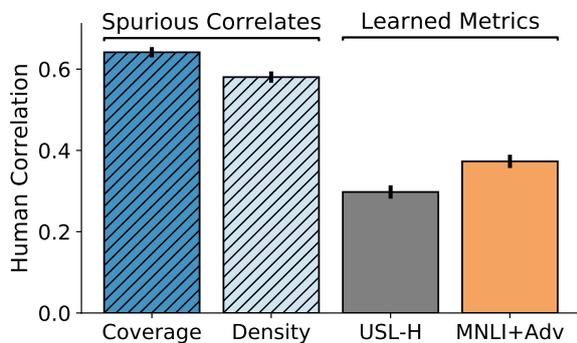

Figure 3: Correlation of the spurious correlates and learned metrics with human scores on groundedness evaluation. Both coverage and density get significantly higher correlations with human scores than the learned metrics.

Metric	Human	Coverage	Density
USL-H	0.298	0.467	**0.515**
MNLI+Adv	0.373	0.451	**0.514**

Table 3: Correlation of USL-H and MNLI+Adv scores with humans vs coverage and density. Both learned metrics have a significantly higher correlation with density than human scores.

For groundedness evaluation[3], both *coverage* and *density* achieve significantly higher correlation with human scores than MNLI+Ad and USL-H. Furthermore, MNLI+Ad and USL-H get a higher correlation with these spurious correlates than human scores (Figure 3).

Despite relatively high correlations on their original datasets, these metrics seem to perform similarly to simple spurious correlations on other datasets. In order to better understand the effectiveness of these reference-free evaluation metrics, we suggest that future research includes comparisons to potential spurious correlates and that research communities come up with a set of potential standard spurious correlates.

4 Learned Metrics in System-level Evaluation

4.1 Pairwise Ranking of Systems

Our example-level analysis demonstrates that recently proposed learned evaluation metrics achieve worse correlations with human scores than spurious correlates for almost all the settings. Since an important goal of building these metrics is to be able to rank arbitrary systems, we analyze whether these concerns we observe at the example level manifest into harms at the system level (i.e., ranking systems incorrectly). In order to study this, we need a large collection of human evaluation data across a wide range of systems. Fabbri et al. (2020) have recently released human evaluations for faithfulness across 16 summarization systems on CNN/DM. Therefore, we focus on system-level rankings of faithfulness for the remainder of the paper.

We first measure pairwise ranking accuracy for all the systems shown in Figure 4.[4] We find that system-level rankings suffer from a similar issue as the example level correlations: density and cover-

---

[3]We do not include MAUDE and DialogRPT results for this task since they perform significantly worse.

[4]Citations corresponding to these systems are included in Appendix A.
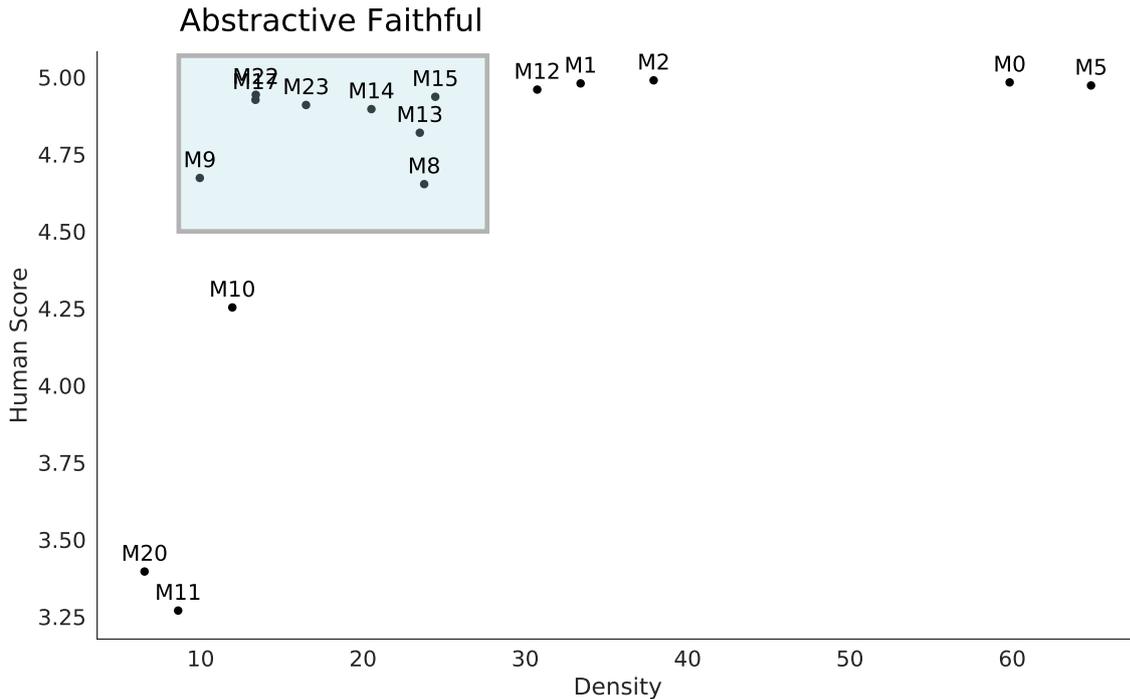
Figure 4: Density and human scores for summarization systems. We analyze the accuracy of the metrics in ranking all the systems vs. ranking the systems within abstractive faithful group, shown in the blue box. Abstractive faithful systems have faithfulness score higher than 4.5 (out of 5) and density lower than 30.

	All Pairs	Within AF
Coverage	56.54	26.60
Density	**81.01**	**40.45**
FactCC	78.87	38.26
DAE	80.39	37.88

Table 4: Accuracy of pairwise ranking across all the systems and within Abstractive Faithful (AF). We observe that the ranking accuracy of all metrics is significantly lower for systems within AF compared to all pairs. Density performs as well as the best learned metric (DAE) in both cases.

age appear as spurious correlations (Table 4). From this observation, we perform a finer-grained analysis and show that these factuality metrics fail on the most important subset of model comparisons: abstractive but faithful summarization system (AF) – where the current state-of-the-art abstractive summarization systems fall.

4.2 Results

Both faithfulness metrics perform relatively well when we look at pairwise ranking accuracy across all pairs of models (Table 4). However, they are unable to improve over *density*, which achieves the highest overall accuracy. When we look at ranking within the abstractive faithful group, we see *density* is no longer a good measure for the faithfulness of a system since these systems are relatively close in terms of density. Similarly, the performance of the learned metrics drops significantly, which is an expected result since our analysis in § 3.1 showed that both FactCC and DAE are spuriously correlated with density. We claim that our system-level analysis is further evidence that these metrics may be relying heavily on simple spurious measures such as word overlap.

These results highlight the importance of performing analyses across different distributions of systems. If we were looking at just the overall ranking accuracy of the metrics, we would conclude that DAE and FactCC correctly measure faithfulness. However, on closer examination, we see that both metrics perform relatively poorly in ranking AF systems, which is arguably the most crucial group since most state-of-the-art systems operate in this regime, and there is substantial interest in building abstractive and faithful summarization systems.
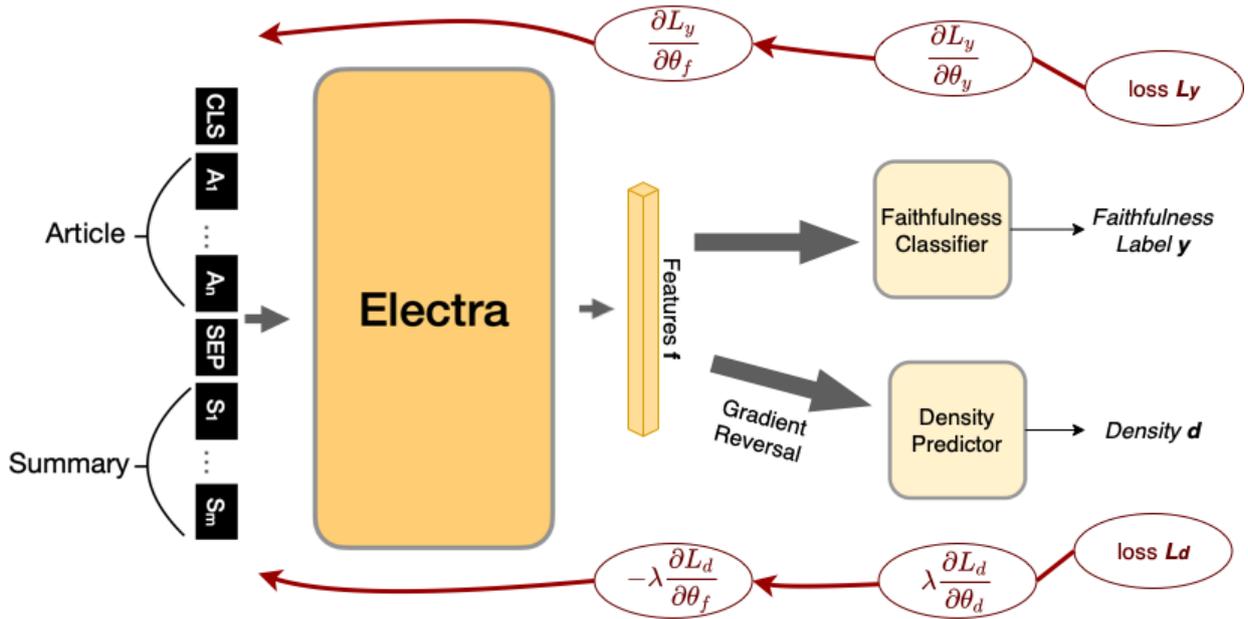
1448

Figure 5: Architecture of adversarial model. The input sequence is first encoded via a pre-trained Electra model, and the representation is used for both faithfulness classification and density prediction. Gradients from the density predictor are reversed in order to make updates to the encoder's parameters, forcing the model to learn representations that are not predicitve of density.

	All Pairs	Within AF
FactCC-Electra	77.85	27.70
FactCC	78.87	38.26
DAE	80.39	37.88
Adversarial	**85.27**	**59.20**

Table 5: Pairwise ranking accuracy for systems across All Pairs vs. Within Abstractive Faithful (AF) for DAE and Adversarial. Adversarially trained metric performs significantly better for the systems within AF than previously proposed metrics.

5 Adversarial Model

In our earlier example-level analysis, we found that learned metrics have higher correlation with spurious correlates than human judgment. We further saw in our system-level analysis that learned metrics for faithfulness are unable to outperform density. One natural question that follows is whether we can build metrics that do well at the systems level by learning representations that rely less on spurious correlates.

In order to do this, we train an entailment based model using the synthetically generated data from FactCC in an adversarial setup similar to Ganin et al. (2016). In particular, our approach augments the standard faithfulness predictor with a density predictor that tries to predict the density of the summary from the model's internal representation. We use this density predictor as an adversary, and our goal is to predict faithfulness while ensuring that it is difficult to predict density using this same representation. To achieve this, the gradients from the density predictor are reversed, which makes it harder to predict the density from the encoder's representation, and thus makes the faithfulness predictions less reliant on density. The model architecture is shown in Figure 5. We initialize the parameter λ to 0 and gradually increase it to 1, following the schedule detailed in Ganin et al. (2016).

We fine-tune a pre-trained Electra model (Clark et al., 2020) using the transformers library (Wolf et al., 2020) for this task. We chose Electra in order to match the model architecture in DAE. Since the original FactCC metric was fine-tuned on BERT, we also fine-tune our own version of FactCC on Electra (FactCC-Electra) as an ablation. Our adversarially trained model is essentially the same as FactCC-Electra, but with an additional adversarial head for predicting density.

Results. We note that the FactCC-Electra model performs worse than the original FactCC, which is consistent with the findings in Goyal and Durrett (2021). Our adversarially trained metric has a significantly lower example-level correlation with density (27.71%), as compared to FactCC (59.10%)

and DAE (76.37%). We find that the adversarial model[5] can achieve a significantly better performance than existing learned evaluation metrics in ranking systems within the abstractive faithful (AF) group (Table 5). This suggests that it is possible to learn effective metrics that are not overly reliant on spurious correlates. Furthermore, our metric is also effective in overall pairwise ranking of the systems achieving 85.27% accuracy.

6 Related Work

Most existing work on assessing the evaluation methodology of evaluation metrics has focused on reference-based evaluation. For example, Mathur et al. (2020) take a critical look at the use of example-level correlations to measure reference-based evaluation metrics in Machine Translation. They show that evaluating these metrics using example-level correlations can be sensitive to the presence of outliers which can lead to false conclusions about a metric's efficacy. Furthermore, Kocmi et al. (2021) show that proper assessment of evaluation metrics is crucial as uninformed use of automated metrics such as BLEU can lead to bad deployment decisions. Caglayan et al. (2020) has shown that automated reference-based evaluation metrics have robustness issues which can cause them to score generated outputs higher than human written outputs. Furthermore, Bhandari et al. (2020) has studied the limitations of reference-based evaluation metrics of text summarization, comparing these metrics across different datasets and application scenarios. In contrast, our work focuses on analyzing learned, reference-free evaluation metrics in summarization and dialog generation, accounting for potential spurious correlates for these evaluation tasks.

There has been some recent work comparing existing reference-free evaluation metrics for text summarization and dialog generation. Pagnoni et al. (2021) has measured the efficacy of existing reference-free faithfulness evaluation metrics of summarization on two different summarization datasets relying on example-level correlations. Similarly, Gehrmann et al. (2021) has evaluated automated metrics of text summarization across a wide range of datasets. Gabriel et al. (2021) has proposed a meta-evaluation framework to evaluate the evaluation metrics looking at certain aspects of

these metrics such as robustness, sensitivity, high correlation with human scores, etc., and measured existing evaluation metrics across these aspects. Yeh et al. (2021) perform a comprehensive study of existing dialog generation metrics across several different datasets and find that the performance of metrics varies widely across datasets.

Gabriel et al. (2021) and Yeh et al. (2021) are the most related to our work since they study robustness of these metrics looking at their performance across different datasets. In our work, however, we explicitly study spurious correlations and show that these may potentially be contributing to the robustness issues. We further present initial promising results suggesting that controlling for these spurious correlates may result in more robust evaluation metrics.

7 Conclusion

In conclusion, we study reference-free evaluation metrics for summarization and dialog generation and show that simply looking at overall example-level correlation with human judgment paints an incomplete picture of the effectiveness of a metric. In particular, we show that these metrics are unable to do better than simple spurious correlates for the task. We see that this trend carries over in system-level ranking for summarization systems, where a spurious correlate for the task performs as well as existing learned evaluation metrics. We find that despite the relatively high overall system-level ranking performance, the learned metrics are not robust to distribution shifts. We show that they fail to properly rank abstractive and (relatively) faithful systems, which is where the current state of the art operates. Finally, we train a faithfulness metric that scores the faithfulness of a summary without relying on the spurious overlap correlate. We show that our metric is more robust across distribution shifts and does better at ranking abstractive, faithful summarization systems.

We suggest that future work in designing reference-free evaluation metrics should be mindful of the distribution of the evaluation data. In particular, metrics should be assessed across different distributions of systems in order to test for robustness and failure modes. Simple spurious correlates can be used as a tool to indicate potential overestimates of the effectiveness of proposed metrics. Finally, we highlight the importance of collecting large-scale human evaluation datasets across a wide

---

range of systems, similar to Fabbri et al. (2020), to enable more comprehensive analyses of evaluation metrics.

8 Acknowledgements

References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages

3739–3748, Brussels, Belgium. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The BEGIN benchmark. *CoRR*, abs/2105.00071.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

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.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van

Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

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

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia. Association for Computational Linguistics.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141, Melbourne, Australia. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 2018. Closed-book training to improve summarization encoder memory. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4067–4077, Brussels, Belgium. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *CoRR*, abs/2107.10821.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen R. McKeown. 2021. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. *CoRR*, abs/2108.13684.

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.

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.

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.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653, New Orleans, Louisiana. 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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. 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.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

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

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5602–5609. AAAI Press.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *ArXiv*, abs/1912.08777.

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.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A   Text Summarization Models

Model Name	Paper
M0	Lead-3 baseline
M1	Zhou et al. (2018)
M2	Dong et al. (2018)
M5	Wu and Hu (2018)
M8	See et al. (2017)
M9	Chen and Bansal (2018)
M10	Gehrmann et al. (2018)
M11	Kryściński et al. (2018)
M12	Hsu et al. (2018)
M13	Pasunuru and Bansal (2018)
M14	Guo et al. (2018)
M15	Jiang and Bansal (2018)
M17	Raffel et al. (2019)
M20	Ziegler et al. (2019)
M22	Lewis et al. (2020)
M23	Zhang et al. (2020)

Table 6: Models that are used in § 4.