Assessing Inter-metric Correlation for Multi-document Summarization Evaluation

Michael Ridenour, Ameeta Agrawal, Olubusayo Olabisi
Portland State University
{mride2,ameeta,oolabisi}@pdx.edu

Abstract

Recent advances in automatic text summarization have contemporaneously been accompanied by a great deal of new metrics of automatic evaluation. This in turn has inspired recent work to re-assess these evaluation metrics to see how well they correlate with each other as well as with human evaluation, mostly focusing on single-document summarization (SDS) paradigm. Although many of these metrics are typically also used for evaluating multi-document summarization (MDS) tasks, so far, little attention has been paid to studying them under such a distinct scenario. To address this gap, we present a systematic analysis of the inter-metric correlations for MDS tasks, while comparing and contrasting the results with SDS models. Using datasets from a wide range of domains (news, peer reviews, tweets, dialogues), we thus study a unified set of metrics under both the task setups. Our empirical analysis suggests that while most reference-based metrics show fairly similar trends across both multi- and single-document summarization, there is a notable lack of correlation between reference-free metrics in multi-document summarization tasks.

1 Introduction

Summarization systems, which aim to preserve salient information from the source text in a more concise form, are being applied to an increasingly diverse range of domains, such as summarizing news articles, messenger-style text conversations, tweets, and so on (Nallapati et al., 2016; Nguyen et al., 2018; Gliwa et al., 2019). Evaluating the performance of these systems is still challenging, and since human evaluation is expensive to obtain, automatic evaluation metrics continue to provide an effective way of evaluating summary quality.

Since no single metric can comprehensively measure every aspect of a summary, it is becoming increasingly common to report system performance in terms of multiple metrics (Fabbri et al., 2021b). As such, it becomes desirable to find a small set of metrics that each reflect different aspects of system performance without redundantly repeating information. Conversely, if a metric is highly correlated with another metric but outperforms it when compared with human evaluation, then that performance difference is more significant (Graham, 2015; Bhandari et al., 2020; Pagnoni et al., 2021). However, in order to do this, one must first understand how these different metrics correlate with each other.

Previous work has focused on studying these metrics under the single-document summarization (SDS) setup, especially news (Bhandari et al., 2020; Fabbri et al., 2021b). However, it is well known that news summarization datasets contain a strong sentence position bias where the most salient information tends to be at the beginning of the article (Nenkova, 2005), which has been shown to have a strong impact on the behavior and performance of some summarization systems (Kryscinski et al., 2019), but does not hold in other domains (Kedzie et al., 2018). Evaluation metrics have also been re-evaluated in the context of scientific articles (Cohan and Goharian, 2016), and more recently, dialogues (Gao and Wan, 2022), both using single documents as input.

In contrast to SDS, multi-document summarization (MDS) is the task of generating a summary from several related documents (Li et al., 2020; Pasunuru et al., 2021; Xiao et al., 2022). Understanding how these metrics estimate MDS tasks, however, remains unexplored. This is notable because many reference-free metrics in particular rely on the source document to evaluate the summary, and when the source consists of stylistically diverse multiple documents, we postulate that it makes the task especially challenging for reference-free metrics. It is unclear whether the automatic evaluation metrics will correlate with each other in the same

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 428 - 438
December 7, 2022 ©2022 Association for Computational Linguistics

way in MDS as they do in SDS tasks. To address these gaps, we present a systematic study on assessing the inter-metric correlations between evaluation metrics for multi-document summarization. Our findings suggest a striking lack of correlation between the reference-free metrics under the MDS paradigm.

Our contributions include the following: (1) We conduct a comprehensive set of experiments for multi-document summarization using several summarization models and datasets from different domains and evaluate them over a *unified set of 16 metrics*; (2) We contextualize our results by drawing comparable insights under the single-document summarization paradigm. (3) Lastly, we summarize our key takeaways and discuss some potential implications of our findings.

2 Related Work

Conventionally, automatic metrics for evaluating summarization systems are mostly reference-based which require human-written reference summaries against which system-summaries can be compared (Lin, 2004; Banerjee and Lavie, 2005). However, since human annotation remains expensive to obtain, automatic evaluation metrics that rely on the source document(s) rather than human-generated reference summaries are becoming increasingly popular (Vasilyev et al., 2020; Scialom et al., 2021).

In parallel to this, researchers have re-assessed how effective these different types of evaluation metrics are, with almost all prior work focused on the single-document framework. Cohan and Goharian (2016) find that ROUGE is not effective at evaluating the performance of summarization systems in the domain of scientific articles. More recently, Bhandari et al. (2020) collect human pyramid-score evaluations (Nenkova and Passonneau, 2004) of sets of 100 summaries generated from 25 top-scoring summarization systems on the CNN/DailyMail dataset (Hermann et al., 2015; Nallapati et al., 2016). They then assess how well 8 different automatic evaluation metrics correlate with the human annotations using the William's test (Williams, 1959), and they also see how well these metrics perform on the shared tasks from the Text Analysis Conferences (TAC). Their analysis finds that most of the metrics fail to generalize well to all the datasets they tested, and that different metrics perform well on different datasets: MoverScore (Zhao et al., 2019) is found to correlate well with human evaluation on TAC-2008, Jensen-Shannon divergence on TAC-2009, and ROUGE-2 on CNN/DM. Similarly, Fabbri et al. (2021b) collect human Likert ratings of 16 systems summarizing 100 documents from CNN/DailyMail, and then use this to assess 14 evaluation metrics. They also find that reference-free metrics are loosely correlated with other metrics. The most recent work is by Gao and Wan (2022) that assesses 18 metrics on 14 systems, generating summaries from the SAMSum dataset (Gliwa et al., 2019) which comprises of messenger-style text conversations.

Type	Dataset	Domain	#Docs/Input
MDS	Multi-News	news	∼2.75
	PeerSum v2	peer reviews	∼7.75
	TSix	tweets	∼35.7
SDS	CNN/DM	news	1
	SAMSum	dialogues	1

Table 1: Statistics of summarization datasets

We also collect system summaries and evaluate them with automatic metrics in our work, except we focus on the correlation between metrics, rather than comparing with human evaluation which is infamously difficult (Gehrmann et al., 2022). While prior work has focused on SDS, our analysis considers both MDS and SDS frameworks, a first such study to our knowledge, across datasets from four different domains.

3 Experimental Setup

3.1 Data

For our experiments, we use three MDS datasets: Multi-News dataset from the news domain (Fabbri et al., 2019), PeerSum which involves summarizing peer reviews of scientific publications (Li et al., 2022), and TSix dataset from the tweets domain (Nguyen et al., 2018). While the first two contain abstractive summaries, the third one contains extractive summaries. Some sample instances from the datasets are included in Appendix A.

As comparison, we also include two abstractive SDS datasets: CNN/DM from the news domain (Hermann et al., 2015), and SAMSum which involves summarizing chat dialogues (Gliwa et al., 2019). Table 1 presents statistics of the five summarization datasets.

3.2 Metrics

In reference-based evaluation, the system-generated summaries are compared to human-written reference summaries, while in unsupervised reference-free evaluation, the system summaries are evaluated using the input source document(s) without relying on human annotations. In this work, we consider a total of 16 widely reported evaluation metrics, 8 each from the reference-based (RB) and reference-free (RF) categories of metrics, which we further group as follows:

1. (RB) Metrics that measure n-gram overlap between the system summary and reference summary: **BLEU**[1] (Papineni et al., 2002), **ROUGE**[2] (Lin, 2004), **METEOR** (Banerjee and Lavie, 2005).

2. (RB) Metrics that use static word embeddings to compare the system and reference summaries: **Embedding Average** (Landauer and Dumais, 1997), **Greedy Matching** (Rus and Lintean, 2012), **Vector Extrema** (Forgues et al., 2014).

3. (RB) Metrics that use contextual representations to compare the system and reference summaries: **MoverScore**[3] (Zhao et al., 2019), **BERTScore**[4] (Zhang* et al., 2020).

4. (RF) Metrics that directly compare the system summary and source document: **Jensen-Shannon divergence**[5] (Lin et al., 2006), **BLANC**[6] (Vasilyev et al., 2020), **SUPERT**[7] (Gao et al., 2020), and **ESTIME**[8] (Vasilyev and Bohannon, 2021).

5. (RF) Metrics that use question-answering to compare the system summary and source document: **SummaQA** (Scialom et al., 2019), **QuestEval**[9] (Scialom et al., 2021).

6. (RF) Metrics that use text generation to measure the conditional probability of generating the summary given the source document, or vice versa: **BARTScore**[10] (Yuan et al., 2021), **Information Difference** (Egan et al., 2021).

3.3 Models

For generating *extractive* summaries, we use **Lead**, **LexPageRank** (Erkan and Radev, 2004), **TextRank** (Mihalcea and Tarau, 2004), **Cluster-CMRW** (Wan and Yang, 2008), **BERT-Ext** and **Longformer-Ext** (Miller, 2019). For generating *abstractive* summaries, we use **BART** (Lewis et al., 2019), **T5** (Raffel et al., 2019), **LED (Longformer Encoder-Decoder)** (Beltagy et al., 2020), and **Pegasus** (Zhang et al., 2020).

In our experiments on the Multi-News dataset (Fabbri et al., 2019), we use a combination of extractive and abstractive models because both types of models were used in the original paper. For comparable results, for the CNN/DM (Hermann et al., 2015) and SAMSum (Gliwa et al., 2019) datasets, we use the model outputs from the SummEval (Fabbri et al., 2021b) and DialSummEval (Gao and Wan, 2022) collections of system summaries, respectively, rather than generating summaries from scratch. Detailed descriptions of these models and the system outputs are included in Appendix B.

3.4 Correlation Analysis

With each dataset we collect system summaries for a set of 100 randomly selected samples from the test set, following recent work on measuring correlations between metrics (Bhandari et al., 2020; Fabbri et al., 2021b; Gao and Wan, 2022). For each sample d_i, $i \in \{1...N\}$ in a dataset $\mathcal{D}$ we generate J summaries from J models, and we denote each summary as s_{ij}, $j \in \{1...J\}$. We use Pearson's r to compute the system-level correlation between two metrics m_1 and m_2 as follows:

$$r^{sys}_{m_1 m_2} = r([\frac{1}{N}\sum_{i=1}^{N} m_1(s_{i1}), ..., \frac{1}{N}\sum_{i=1}^{N} m_1(s_{iJ})], [\frac{1}{N}\sum_{i=1}^{N} m_2(s_{i1}), ..., \frac{1}{N}\sum_{i=1}^{N} m_2(s_{iJ})]).$$

[1] https://github.com/Maluuba/nlg-eval is used for BLEU, METEOR, and the word embedding-based metrics

[2] https://github.com/Diego999/py-rouge

[3] https://github.com/AIPHES/emnlp19-moverscore

[4] https://github.com/Tiiiger/bert_score

[5] github.com/UKPLab/coling2016-genetic-swarm-MDS

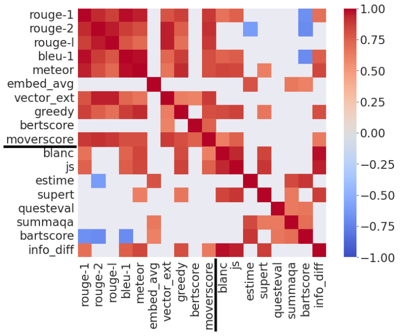
[6] BLANC-tune, which uses the summary to first fine-tune the model

[7] https://github.com/Yale-LILY/SummEval is used for SUPERT and SummaQA

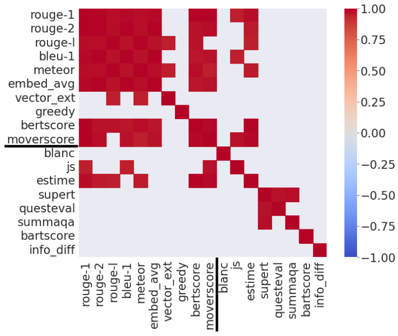
[8] https://github.com/PrimerAI/blanc is used for ESTIME, BLANC, and Information Difference

[9] https://github.com/ThomasScialom/QuestEval

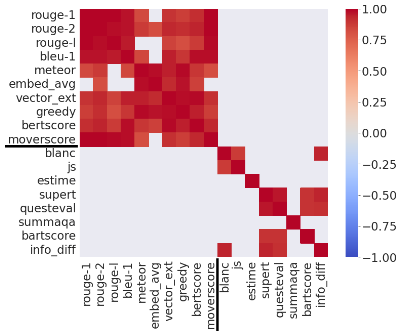
[10] https://github.com/neulab/BARTScore is used for BARTScore (source -> hypothesis)



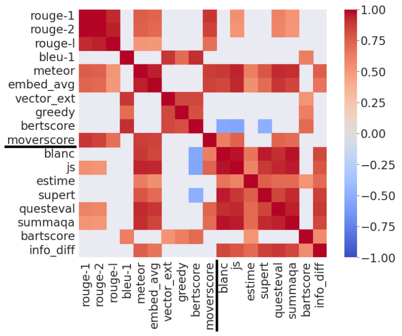
(a) Multi-News



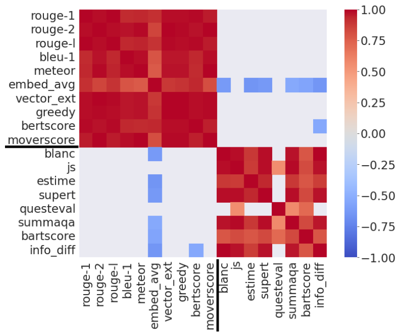
(b) PeerSum



(c) TSix



(d) CNN/DM



(e) SAMSum

Figure 1: Pearson's r correlation between metrics on the system level for the MDS datasets in the top row – (a) Multi-News, (b) PeerSum, and (c) TSix, followed by the SDS datasets in the bottom row – (d) CNN/DM, and (e) SAMSum. Note that only statistically significant correlations are displayed ($p \leq 0.05$), and reference-based and reference-free metrics are delineated by a line.

4 Results and Analysis

In this section, we discuss the results of two main experiments where we investigate the inter-metric correlations across two types of summarization (multi-document and single-document) over four different domains (peer reviews, tweets, news, and dialogues). In each experiment we calculate the Pearson's r correlations between metrics and report statistically significant values ($p \leq 0.05$).

4.1 Multi-document summarization

Figures 1a, 1b, and 1c present the results of correlation analysis on the Multi-News, PeerSum, and TSix multi-document summarization datasets. Across all three datasets the reference-based metrics correlate positively with each other, whereas correlations within the reference-free metrics are noticeably fragmented, with PeerSum exhibiting the most fragmentation. This is likely due to the higher diversity in the source documents that is intrinsic to these MDS tasks, especially in PeerSum where roughly 9% of ICLR paper reviews have a rating difference ≥ 5 (Li et al., 2022). This makes it harder to compare the source documents to the summary in a consistent manner. Moreover, between the two broad categories of metrics, reference-based and reference-free, no consistent correlation can be observed.

4.2 Single-document summarization

Figures 1d and 1e present the results of evaluating single-document summarization datasets (CNN/DM and SAMSum, respectively) on the same set of metrics as used in the previous section for a comparable discussion. In contrast to the observations made on the MDS datasets, here we see a strong positive correlation within almost all reference-free metrics, on both the datasets. Futhermore, it is easy to see, especially on SAMSum dataset, that reference-based and reference-free metrics are highly correlated to each other within their respective groups, but there is little positive correlation between groups (we see some statistically significant anti-correlation), confirming the results found in Gao and Wan (2022). On CNN/DM, although the results appear to be a bit more mixed, clusters of high correlation within fine-grained categories of evaluation metrics are clearly observed – metrics based on static or contextual representations (Vector Extrema, Greedy

Matching, BERTScore), metrics that use question-answering or other means to compare the system summary and source document (QuestEval, SummaQA, BLANC, Jensen-Shannon, ESTIME, SUPERT), and the metrics that use text generation (BARTScore and Information Difference) are all strongly correlated.

4.3 Discussion

In comparing all the results of Figure 1, several observations are made, thus allowing us to put forward some recommendations.

- **Reference-based vs. Reference-free metrics**. First, given almost no agreement between reference-based and reference-free metrics, it appears that these families of metrics measure distinct qualities of a summary, suggesting the need for reporting some metrics from each category, regardless of the summarization framework or dataset domain.

- **Domain-based observations**. Most noticeably, both the datasets from the news domain, whether MDS (Multi-News) or SDS (CNN/DM), exhibit similar and arguably more fragmented heatmaps. This is in sharp contrast to the results from the other three domains (peer reviews, tweets, and dialogues), all of which show similar trends. This indicates that conclusions drawn for these evaluation metrics under one domain may not hold true for another. Thus it is important to consider the differences in domain while introducing and re-assessing evaluation metrics.

- **Similarities between MDS and SDS analysis**. Across both paradigms of MDS and SDS, the reference-based metrics tend to behave similarly, i.e., correlate significantly positively with each other (with CNN/DM being somewhat of an exception).

- **Differences between MDS and SDS analysis**. In SDS tasks, in general reference-free metrics tend to show high correlation with each other suggesting that reporting a small subset of them might be adequate. However, rather interestingly, *in the case of MDS datasets, the reference-free metrics indicate little to no correlation*. We hypothesize that this is likely due to the unique construction of multiple source documents being so diverse.

  The striking differences between the behavior of reference-free metrics under SDS and MDS paradigms, therefore, motivate the need for further investigation into how reference-free metrics are applied to MDS tasks.

5 Conclusions and Future Work

We conduct an in-depth assessment of the correlations between numerous evaluation metrics, including those that use reference summaries and those that do not, in the context of multi-document summarization tasks. As a further investigation, we also evaluate single-document summarization datasets on the same set of metrics. Our results indicate that evaluation metrics behave noticeably differently when studied under MDS and SDS paradigms, which makes metrics for MDS an interesting avenue of research to be explored further. Moreover, measuring how these metrics correlate with different dimensions of human evaluation on MDS might be beneficial.

Limitations

As has been recently pointed out in Deutsch et al. (2022), using system outputs on the full test set rather than just 100 samples can make these results much more robust by giving a lower-variance estimate of the inter-metric correlations.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 4106–4118, Online. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

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.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. *arXiv preprint arXiv:1604.00400*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Bandit-Sum: 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.

Nicholas Egan, Oleg V. Vasilyev, and John Bohannon. 2021. Play the shannon game with language models: A human-free approach to summary evaluation. *ArXiv*, abs/2103.10918.

Güneş Erkan and Dragomir R. Radev. 2004. LexPageRank: Prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 365–371, Barcelona, Spain. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021a. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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 6866–6880, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. 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 1479–1491, Online. Association for Computational Linguistics.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

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.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. 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.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. 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.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2022. Peersum: A peer review dataset for abstractive multi-document summarization.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. *arXiv preprint arXiv:2005.10043*.

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

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 463–470, New York City, USA. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *AAAI*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Minh-Tien Nguyen, Dac Viet Lai, Huy-Tien Nguyen, and Le-Minh Nguyen. 2018. TSix: A human-involved-creation dataset for tweet summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

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.

Ramakanth Pasunuru and Mohit Bansal. 2019. Continual and multi-task architecture search. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1911–1922, Florence, Italy. Association for Computational Linguistics.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779.

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. *arXiv preprint arXiv:1910.10683*.

Vasile Rus and Mihai Lintean. 2012. An optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems*, ITS'12, page 675–676, Berlin, Heidelberg. Springer-Verlag.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.

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.

Oleg Vasilyev and John Bohannon. 2021. ESTIME: Estimation of summary-to-text inconsistency by mismatched embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 94–103, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

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.

Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 299–306, New York, NY, USA. Association for Computing Machinery.

Evan James Williams. 1959. *Regression Analysis*. Wiley, New York.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122, Online. Association for Computational Linguistics.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *AAAI*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. 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.

A Dataset Samples

Table 2 presents a sample instance of input documents and corresponding reference summary from the Multi-News dataset, Table 3 presents a sample from the PeerSum dataset, and Table 4 presents a sample from the TSix dataset. Reference-free metrics used the full source documents (no truncation) for evaluation.

B Model Details

B.1 Multi-News dataset

We generate summaries with BART (Lewis et al., 2019), T5 (Raffel et al., 2019), LED (Beltagy et al., 2020), Pegasus (Zhang et al., 2020), and Longformer (Beltagy et al., 2020). Additionally, we used the system outputs provided by (Fabbri et al., 2019), which includes LexPageRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004), MMR (Fabbri et al., 2019), Transformer (Vaswani et al., 2017), PG-BRNN (See et al., 2017), and Hi-MAP (Fabbri et al., 2019). [11]

B.2 PeerSum dataset

For generating *abstractive* summaries, we use four neural-based abstractive summarization systems. We concatenate the input documents. All pre-trained model checkpoints accessed from the Huggingface library (Wolf et al., 2019) were further fine-tuned on PeerSum dataset (Li et al., 2022), except for Pegasus. The systems include BART (Lewis et al., 2019) which combines a bidirectional encoder with an auto-regressive decoder, T5 (Raffel et al., 2019) which is an encoder-decoder model trained using teacher forcing, LED (Longformer Encoder-Decoder) (Beltagy et al., 2020) which is a variant of the Longformer model with both encoder and decoder transformer stacks, also scaling linearly with the input, and Pegasus (Zhang et al., 2020) which is a sequence-to-sequence model with gap-sentences generation as a pretraining objective. The system outputs we use in our experiments were generated from 100 samples from the test set. Reviews, comments, and replies were used to generate the summaries.

B.3 TSix dataset

For generating *extractive* summaries, we use systems representing a mixture of traditional methods and state-of-the-art neural-based architectures.

[11] `https://github.com/Alex-Fabbri/Multi-News`

Our models consist of Lead[12] which extracts the first n-tweets, LexPageRank (Erkan and Radev, 2004) and TextRank[13] (Mihalcea and Tarau, 2004) which are unsupervised graph-based ranking methods, ClusterCMRW (Wan and Yang, 2008), BERT-Ext (Miller, 2019), an extractive summarization model[14] built on top of BERT (Devlin et al., 2018) which uses K-means clustering to select sentences closest to the centroid as the summaries, and similarly, Longformer-Ext which uses embeddings from the pretrained Longformer model (Beltagy et al., 2020). The 100 system outputs we use in our experiments are roughly 15 tweets long on average and were generated from samples that have between 50-100 tweets as input.

B.4 CNN/DM dataset

For the CNN/DM dataset (Hermann et al., 2015), we used the system outputs provided by (Fabbri et al., 2021b). This consists of 16 models, each with 100 outputs.[15]

Models: LEAD-3, NEUSUM (Zhou et al., 2018), BanditSum (Dong et al., 2018), RNES (Wu and Hu, 2018), Pointer-generator (See et al., 2017), Fast-abs-rl (Chen and Bansal, 2018), Bottom-Up (Gehrmann et al., 2018), Improve-abs (Kryściński et al., 2018), Unified-ext-abs (Hsu et al., 2018), ROUGESal (Pasunuru and Bansal, 2019), Multi-task (Ent+QG) (Guo et al., 2018), T5 (Raffel et al., 2019), GPT-2 (zero-shot) (Radford et al., 2019), BART (Lewis et al., 2019), Pegasus (C4) and Pegasus (dynamic mix) (Zhang et al., 2020).

B.5 SAMSum dataset

For the SAMSum dataset (Gliwa et al., 2019), we used system outputs provided by (Gao and Wan, 2022). This consists of 14 models, each with 100 outputs.[16] The dataset includes the human-written reference and two extractive models in the system outputs; excluding these increases correlation between reference-free and reference-based metrics but does not significantly change correlations within those groups.

Models: LEAD-3, LONGEST-3, Pointer-generator (See et al., 2017), Transformer (Vaswani et al., 2017), BART (Lewis et al., 2019), Pegasus (Zhang et al., 2020), UniLM (Dong et al., 2019), CODS (Wu et al., 2021), ConvoSumm (Fabbri et al., 2021a), MV-BART (Chen and Yang, 2020), PLM-BART (Feng et al., 2021), Ctrl-DiaSumm (Chen et al., 2021), S-BART (Chen and Yang, 2021).

[12] `https://github.com/PKULCWM/PKUSUMSUM` is used for Lead, LexPageRank, and ClusterCMRW

[13] `https://github.com/RaRe-Technologies/gensim`

[14] `https://pypi.org/project/bert-extractive-summarizer/`

[15] `https://github.com/Yale-LILY/SummEval`

[16] `https://github.com/kite99520/DialSummEval`

Input Documents (News Articles)
d_1: after a year in which liberals scored impressive, high-profile supreme court victories, conservatives could be in line for wins on some of this term's most contentious issues, as the justices consider cases that could gut public sector labor unions and roll back affirmative action at state universities. however, as the court's new term kicks off monday, uncertainty surrounds several other politically potent cases that could wind up on the court's agenda ... d_2: the new term's biggest rulings will land in june, as the 2016 presidential campaign enters its final stretch, and they will help shape the political debate. "constitutional law and politics are certainly not the same thing, but they are interrelated, never more so than in a presidential election year that will likely determine who gets to appoint the next justice or two or three, " said vikram d. amar, dean of the university of illinois college of law... d_3: the death penalty is shaping up to be a big issue for the supreme court as it begins a new term monday, with at least six capital-punishment cases on the docket and a recent wave of executions keeping the justices up late to field last-minute appeals. in the weeks ahead , the court is set to hear arguments over the constitutionality of capital sentences in florida, georgia, kansas and pennsylvania...
Reference Summary
the supreme court is facing a docket of high-profile political cases that will test whether recent liberal victories were more fluke or firm conviction, the new york times reports. the court — which is divided 5-4 for conservatives, but saw justice roberts vote liberal on obamacare and same-sex marriage — will look at cases including unions, affirmative action, and possibly abortion...

Table 2: Example instance from Multi-News dataset

Input Documents (Reviews)
$d_1(review)$: This paper proposes a method to train neural networks with low precision. However, it is not clear if this work obtains significant improvements over previous works. Note that: 1) Working with 16bit, one can train neural networks with little to no reduction in performance. For example, on ImageNet with AlexNet one gets 45.11% top-1 error if we don't do anything else, and 42.34% (similar to the 32-bit result) if we additionally adjust the loss scale... $d_2(reply)$: We sincerely appreciate the reviewer for the comments, which indeed helps us to improve the quality of this paper. In our revised manuscript, we keep the last layer in full precision for ImageNet task (both BNN and DoReFa keep the first and the last layer in full precision). Our results have been improved from 53.5/28.6 with 28CC to 51.7/28.0 with 2888 bits setting. Results of other patterns are updated in Table4... ... ... $d_5(review)$: The authors propose WAGE, which discretized weights, activations, gradients, and errors at both training and testing time. By quantization and shifting, SGD training without momentum, and removing the softmax at output layer as well, the model managed to remove all cumbersome computations from every aspect of the model, thus eliminating the need for a floating point unit completely. Moreover, by keeping up to 8-bit accuracy, the model performs even better than previously proposed models. I am eager to see a hardware realization for this method because of its promising results...
Reference Summary (Meta-Review)
High quality paper, appreciated by reviewers, likely to be of substantial interest to the community. It's worth an oral to facilitate a group discussion.

Table 3: Example instance from PeerSum dataset

Input Documents (Tweets)
d_1: Tech company Nanoco says #Brexit could limit supply of talent. d_2: #Pound closes at another 30 year low. Down to $1.21, fallen 7% in 10 days since #TheresaMay's "hard #Brexit" speech. #GBP. . . . d_3: I hope this radio host has a lot of mics, because he keeps dropping them. #brexit. d_4: Today's guest article: Gerald Stubbs laments #Britain losing 40 years of progress because of #Brexit. Please share: htt. . . . d_5: Perhaps we should be pleased and encouraged to see that they're worried and anxious enough about derailment of #Brexit to resor. . . . d_6: How to save what is left of #Greece? Here's one hint: #Brexit.. d_7: Jacob Rees Mogg's 'Ladybird Constitution'. via #Brexit #jacobreesmogg. .... d_{62}: . 'Leaked Treasury papers show UK Government #brexit chaos will damage Scottish economy'. d_{63}: GBPUSD Rallying on Back of Potential Brexit Turn Around. d_{64}: Brexit 'will stunt national living wage growth by 10p an hour'. d_{65}: UK Prime Minister May backs down on parliament vote over her Brexit terms — South China Morning Post.
Reference Summary
Prof Patrick Minford:: EU and trade #EU #brexit #referendum #voteleave 9.. Good Ganeha you think you have an understanding how dim #Brexit vote leave people are... And then you see new evidence. . . .. Now Dutch wants own EU vote & Czechs say they might leave #EU #brexit #referendum #voteleave 4.. Pound Soars as Hard Brexit Fears Recede, US Dollar Aims Higher DailyFX on #GBPUSD.. UK Prime Minister May backs down on parliament vote over her Brexit terms: Prime Minister Theresa May has acc.... IEA cuts oil demand forecast for 2017 #healthinnovations #pharma #banking #stocks #Brexit #oil. . . .. Leave EU and we'll make your lives a misery: Juncker's warning to Britain #EU #brexit #referendum #voteleave 3.. Still would be less crazy than hard Brexit.... . . ..

Table 4: Example instance from TSix dataset