Hiding in Plain Sight: Insights into Abstractive Text Summarization

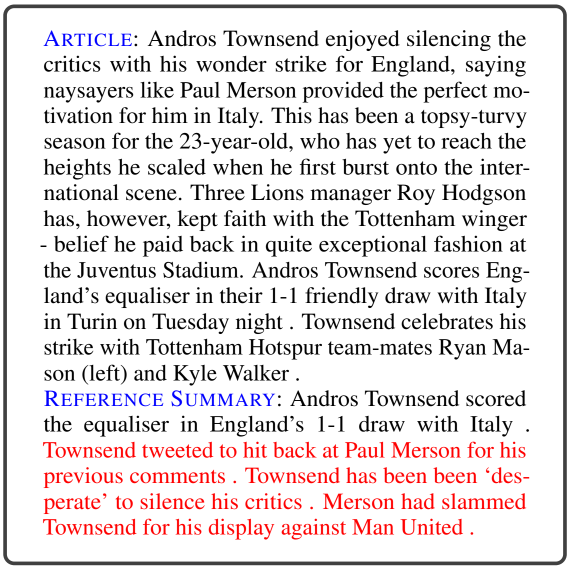
Vivek Srivastava, Savita Bhat, Niranjan Pedanekar
TCS Research
Pune, Maharashtra, India
{srivastava.vivek2, savita.bhat, n.pedanekar}@tcs.com

Abstract

In recent years, there has been growing interest in the field of abstractive text summarization with focused contributions in relevant model architectures, datasets, and evaluation metrics. Despite notable research advances, previous works have identified certain limitations concerning the quality of datasets and the effectiveness of evaluation techniques for generated summaries. In this context, we examine these limitations further with the help of three quality measures, namely, *Information Coverage*, *Entity Hallucination*, and *Summarization Complexity*. As a part of this work, we investigate two widely used datasets (*XSUM* and *CNN-DM*) and three existing models (*BART*, *PEGASUS*, and *BRIO*) and report our findings. Some key insights are: 1) Cumulative ROUGE score is an inappropriate evaluation measure since few high-scoring samples dominate the overall performance, 2) Existing summarization models have limited capability for information coverage and hallucinate to generate factual information, and 3) Compared to the model-generated summaries, the reference summaries have lowest information coverage and highest entity hallucinations reiterating the need of new and better reference summaries.

1 Introduction

Abstractive text summarization (ATS) is the process of compressing given textual content into short and concise form by paraphrasing or rewriting the most important information from the source. Considering the high-level language understanding, reasoning, and generation capabilities required for ATS, considerable improvements are reported in this field with contributions such as large-scale datasets (Gliwa et al., 2019; Ladhak et al., 2020), use of innovative techniques/architectures (Liu and Liu, 2021), and novel evaluation metrics for effective validation. Recently, significant interest has been observed in examining the quality of summarization datasets (Tejaswin et al., 2021), reliability of evaluation metrics (Fabbri et al., 2021), architectural choices, and overall impact of these on model performance. In this paper, we re-evaluate the quality of textual content from summarization datasets and generated summaries with *Information Coverage*, *Entity Hallucination*, and *Summarization Complexity* as primary dimensions of evaluation.



ARTICLE: Andros Townsend enjoyed silencing the critics with his wonder strike for England, saying naysayers like Paul Merson provided the perfect motivation for him in Italy. This has been a topsy-turvy season for the 23-year-old, who has yet to reach the heights he scaled when he first burst onto the international scene. Three Lions manager Roy Hodgson has, however, kept faith with the Tottenham winger - belief he paid back in quite exceptional fashion at the Juventus Stadium. Andros Townsend scores England's equaliser in their 1-1 friendly draw with Italy in Turin on Tuesday night . Townsend celebrates his strike with Tottenham Hotspur team-mates Ryan Mason (left) and Kyle Walker .
REFERENCE SUMMARY: Andros Townsend scored the equaliser in England's 1-1 draw with Italy . Townsend tweeted to hit back at Paul Merson for his previous comments . Townsend has been been 'desperate' to silence his critics . Merson had slammed Townsend for his display against Man United .

Figure 1: Example from the CNN-DM dataset. The highlighted sentences in the reference summary contains facts missing from the source article.

Popular summarization datasets, XSUM (Narayan et al., 2018) and CNN-DM (Hermann et al., 2015; Nallapati et al., 2016) (see Table 1), are known to have major issues such as factual consistency (Maynez et al., 2020; Tam et al., 2022; Laban et al., 2022), low degree of summarization complexity (Tejaswin et al., 2021), and layout biases (Kryściński et al., 2019). Figure 1 shows an example where the reference summary contains the facts that are missing from the source article. The models trained on these datasets tend to pick up these limitations and thus are unreliable for any real-world application.

Among all reference-free (Vasilyev et al., 2020;

The Fourth Workshop on Insights from Negative Results in NLP, pages 67–74
May 5, 2023 ©2023 Association for Computational Linguistics

Dataset	Train/Val/Test	Description
XSUM	204k/11k/11k	BBC news articles (1 sentence summaries)
CNN-DM	287k/13k/11k	CNN & DailyMail news articles (3-4 sentences summaries)

Table 1: ATS datasets overview.

Gao et al., 2020) and reference-dependant (Zhang et al.; Zhao et al., 2019) metrics proposed to date, ROUGE (Lin, 2004) is preferred owing to its ease of interpretation, usage, and comparison with other baselines even though it misses out several quality evaluation dimensions such as factuality and informativeness (Bhandari et al., 2020; Pagnoni et al., 2021; Goyal et al., 2022; Deutsch and Roth, 2021; Akter et al., 2022).

In this paper, we examine two widely used datasets (XSUM and CNN-DM) and analyze the performance of three ATS models (BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), and BRIO (Liu et al., 2022)) on three interpretable quality evaluation dimensions. In contrast to the similar existing works where the human-based evaluation with a very small subset of datasets is considered (Fabbri et al., 2021; Pagnoni et al., 2021), we present a computational framework for these dimensions. We believe that the framework is especially useful in reducing the dependence on human-based evaluation for the quality of the datasets and ATS models.

Model	XSUM			CNN-DM		
	R1	R2	RL	R1	R2	RL
BART	45.14	22.27	37.25	44.16	21.28	40.9
PEGASUS	47.21	24.56	39.25	44.17	21.47	41.11
BRIO	49.07	25.59	40.4	47.78	23.55	44.57

Table 2: Evaluation results on the ROUGE metric.

2 Quality Evaluation Dimensions

In this section, we define three dimensions for quality evaluation. We examine the performance of ATS models over these dimensions. We report the ROUGE-based performance of these models for comparison (see Table 2). We also explore the reference summaries on these dimensions. We denote model-generated zero-shot summary as zs, reference summary as ref, and article as A.

1. **Information coverage**: A high-quality summary highlights the information present in the source document. We explore the information coverage of a summary from two perspectives: topical coverage and key information coverage. In contrast to the naive word overlap between the generated and reference summaries in ROUGE, we consider an informed overlap of the summary with the source article in both formulations.

   Topical coverage (TC): An article usually discusses multiple aspects/topics to present facts and information (see Appendix). The ROUGE-based evaluation fails to measure the topical coverage of the generated summary. To examine this further, we divide the article A into a sequence of topics using the sentence similarity-based topic-segmentation algorithm, C99 (Choi, 2000). We select C99 due to the fast topic segmentation and flexibility to plug and play with different sentence representation models. We use the sentence BERT representations (Reimers and Gurevych, 2019) to segment the article into multiple topics. Each topic contains a sequential list of sentences. We consider a topic T from article A covered by the summary if at least k words from the summary[1] exist in T. Formally,

$$TC(zs, A, k) = 100 * \frac{f_{TC}(zs, A_{topics}, k)}{|A_{topics}|} \quad (1)$$

   where $f_{TC}(.)$ measures the number of topics covered by the summary (constrained by k).

   Key information coverage (KIC): A document summary, by definition, should cover the key information presented in the source document. We identify the key information in the source document using an unsupervised keyphrase extraction tool, YAKE (Campos et al., 2020)[2] (see Appendix). Formally, we define KIC as:

$$KIC(zs, A) = 100 * \frac{f_{KIC}(zs, A_{key-info})}{|A_{key-info}|} \quad (2)$$

   where $f_{KIC}(.)$ measures the number of keyphrases in A that exist in the summary.

2. **Entity hallucination** (EH): In Figure 1 (also see Appendix), we present an example where the summary contains the entities missing from the article A. We consider a model to be entity-hallucinated if it generates an entity missing from the article (Tam et al., 2022). We use an

[1] we preprocess the summary to remove stopwords using the gensim library: `https://github.com/RaRe-Technologies/gensim`

[2] based on our manual analysis, we set ngram-size as 4, dedup-lim as 0.5 and select the key-phrases with a score less than 0.1

18-class named-entity recognition module from spacy[3] to detect the entities. Formally,

$$EH(zs, A) = 100 * \frac{f_{EH}(zs_{entities}, A)}{|zs_{entities}|} \quad (3)$$

where $f_{EH}(.)$ measures the number of entities in the summary that are missing from the article.

3. **Summarization complexity**: We consider summarization complexity to be correlated with the measure of extractiveness in the samples. This complexity could potentially influence the model's performance. For instance, ATS models with a higher tendency to copy text fragments from the source document could achieve high ROUGE scores on samples where the reference summaries are more extractive. We examine this by using a phrase overlap (PO) based formulation. We define phrase overlap between the model-generated summary and the article as:

$$PO_{article}(zs, A, n) = 100 * \frac{|zs_n \cap A_n|}{|zs_n|} \quad (4)$$

Similarly, PO between the model-generated and reference summary is given as:

$$PO_{ref}(zs, ref, n) = 100 * \frac{|zs_n \cap ref_n|}{|zs_n|} \quad (5)$$

Here, zs_n, A_n, and ref_n denote the phrases containing n-tokens in the zero-shot summary, article, and reference summary respectively.

3 Analysis

In this section, we discuss the insights from each of these dimensions. In all our analyses, we divide the samples in the test set of both datasets into four groups. Each group contains 25% samples from the original test set sorted based on the ROUGE-L score. Group 1 (G1) contains samples with the lowest ROUGE-L score whereas group 4 (G4) contains samples with the highest ROUGE-L score. While reporting the results for the reference summary, we use the groups identified using the ROUGE-L ranking of samples with the BRIO model. We report the average scores for each group (see Tables 3 and 4 for information coverage, Table 5 for entity hallucinations, and Tables 6 and 7 for summarization complexity). Some key observations are:

Models trained on the CNN-DM dataset tends to show higher information coverage. This tendency could also be partially attributed to the longer and more extractive summaries generated with the CNN-DM dataset. The gap for topical coverage between both datasets widens further as we increase the value of k.

k	Model	XSUM				CNN-DM			
		G1	G2	G3	G4	G1	G2	G3	G4
1	BART	**76.47**	**81.04**	**81.82**	**81.97**	88.78	91.62	92.22	92.59
	PEG	73.19	79.47	79.68	80.23	87.25	90.78	91.59	91.73
	BRIO	76.96	80.95	81.42	81.34	**91**	**92.58**	**93.02**	**93.51**
	Ref	71.40	78.80	79.95	80.93	86.98	90.71	91.41	92.45
2	BART	54.19	60.24	61.35	**61.17**	79.65	84.23	85.07	85.99
	PEG	50.26	57.88	57.73	58.52	76.87	82.89	84.20	84.56
	BRIO	**54.52**	**60.08**	**59.88**	60.46	**84.02**	**86.40**	**87.47**	**87.94**
	Ref	47.13	56.36	57.55	59.12	75.66	82.14	84.25	85.61
5	BART	**13.31**	**14.91**	**16.03**	**14.80**	54.68	62.13	64.33	66.19
	PEG	11.24	13.16	13.09	12.79	50.86	59.95	62.77	63.58
	BRIO	13.30	14.79	15.33	14.63	**61.97**	**67.98**	**69.07**	**70.03**
	Ref	7.97	11.71	12.30	13.63	45.67	57.50	61.50	65.30

Table 3: Topical coverage on k = 1, 2, and 5. For each k, we highlight minimum TC and **maximum** TC for a group within a dataset. A higher TC is preferred.

Model	XSUM				CNN-DM			
	G1	G2	G3	G4	G1	G2	G3	G4
BART	11.69	**12.08**	11.82	**11.20**	37.55	42.55	44.11	46.54
PEG	**11.81**	11.79	11.15	10.59	33.95	39.18	41.71	43.88
BRIO	11.66	11.97	**11.91**	10.89	**42.36**	**45.75**	**47.63**	**49.62**
Ref	9.60	10.92	10.93	10.51	24.90	30.36	33.60	38.33

Table 4: Key information coverage. We highlight minimum KIC and **maximum** KIC for a group within a dataset. A higher KIC is preferred.

BART gives tough competition to BRIO. Although BRIO gets the highest TC and KIC score on the CNN-DM dataset, BART performs competitively. On the XSUM dataset, both models perform equally well. PEGASUS has the worst TC among all three models suggesting that the generated summaries with PEGASUS are limited in their capability to cover the overall source document.

We need new reference summaries! It is interesting to note that the reference summaries show worst KIC than all three models suggesting that the ATS model's capability to cover key information is limited due to training on these poor-quality reference summaries. Also, the topical coverage of reference summaries is significantly lower in G1 compared to other groups in both datasets, denoting the need for targeted analysis for this group.

Models trained on the XSUM dataset tend to show higher entity hallucination. EH is more prominent in the models trained on the XSUM dataset due to the inherent nature of the dataset (i.e., very high EH score of reference summaries), which calls for the need to look beyond word overlap-based metrics like ROUGE while training and evaluating the ATS models. Also, the high EH of reference summaries in both datasets is

[3] https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.5.0

concerning since it directly limits the capability of proposed techniques for ATS.

The ROUGE-based bench-marking of ATS models is inadequate. In addition to giving tough competition to BRIO for information coverage, BART consistently shows the least EH than the other two models. PEGASUS and BRIO have a similar degree of EH on both datasets (see class-wise EH distribution in Appendix). Low information coverage and high EH of PEGASUS compared to BART contradicts PEGASUS's superior behavior based on the ROUGE score (see Table 2). It reiterates the need for an alternative bench-marking of the ATS models.

Model	XSUM				CNN-DM			
	G1	G2	G3	G4	G1	G2	G3	G4
BART	36.52	**38.80**	**40.95**	**44.96**	**2.08**	**1.38**	**1.29**	**1.37**
PEG	**34.91**	40.36	45.12	48.22	5.87	5.53	5.49	4.91
BRIO	40.27	43.41	45.84	49.52	6.55	4.42	3.99	3.61
Ref	46.01	46.43	50.54	52.61	15.03	12.37	10.68	7.72

Table 5: Entity hallucinations. We highlight maximum EH and **minimum** EH for a group within a dataset. A lower EH is preferred.

Articles in the CNN-DM dataset are easier to summarize? The models trained on the CNN-DM dataset tend to copy text fragments from the source article, and this behavior is more prominent in high ROUGE scoring samples (i.e., G4). BART shows a very-high tendency to copy content from the article and manages to perform well on the ROUGE-based evaluation. It further highlights the extractive nature of the reference summaries in the CNN-DM dataset that guides the model to learn to copy content from the source document.

The tendency to be more abstractive is costly for BRIO! BRIO-generated summaries are more abstractive in nature, especially in the low ROUGE-scoring group G1. The significantly lower PO_{ref} score in this group compared to other groups results in a lower ROUGE score suggesting that the abstractiveness proves costly for BRIO.

ROUGE score is dominated by a few samples. For both the datasets, all models show a sharp increase in the PO_{ref} score as we move from G1 to G4 (see Table 7), suggesting that only a small proportion of samples contribute heavily towards the overall ROUGE score. The gap between the groups widens as we increase the phrase length.

XSUM and CNN-DM datasets are NOT the benchmark datasets for the ATS task. As discussed earlier, the reference summaries in the CNN-DM dataset are more extractive in nature. It is interesting to note that the extractive text summarization models built on this dataset show comparative performance to the ATS models (An et al., 2022). In contrast, the reference summaries in the XSUM dataset are more abstractive with a higher degree of hallucination, making them unsuitable for effective utilization.

n	Model	XSUM				CNN-DM			
		G1	G2	G3	G4	G1	G2	G3	G4
1	BART	64.94	64.72	64.18	62.28	94.74	94.94	94.77	95.16
	PEG	63.92	62.83	61.18	59.69	89.27	89.92	90.12	90.76
	BRIO	62.80	62.31	61.41	59.76	88.03	89.75	90.51	91.92
	Ref	**52.72**	**54.98**	**55.27**	**55.71**	**75.45**	**79.77**	**82.57**	**86.58**
2	BART	23.61	22.38	21.74	20.23	85.38	85.23	84.91	86.02
	PEG	24.80	21.40	19.34	17.93	74.59	74.59	74.84	77.03
	BRIO	20.99	19.79	19.06	17.86	61.72	65.48	68.05	72.92
	Ref	**11.44**	**13.15**	**13.48**	**14.66**	**32.82**	**38.91**	**44.45**	**54.53**
3	BART	9.82	8.13	7.79	7.18	77.40	76.72	76.07	77.76
	PEG	12.35	8.67	6.74	5.95	64.01	63.17	63.33	66.34
	BRIO	6.75	6.18	5.93	5.52	42.74	46.65	49.92	56.96
	Ref	**2.35**	**3.07**	**3.26**	**4.04**	**16.12**	**20.3**	**25.32**	**36.61**

Table 6: Phrase overlap with the article on n = 1, 2, and 3. For each n, we highlight **minimum** $PO_{article}$ and maximum $PO_{article}$ for a group within a dataset. A higher $PO_{article}$ suggests more extractive summaries.

n	Model	XSUM				CNN-DM			
		G1	G2	G3	G4	G1	G2	G3	G4
1	BART	24.99	37.21	46.74	63.28	22.40	32.11	39.17	51.01
	PEG	25.87	39.97	50.13	**68.43**	25.69	35.70	43.24	**56.23**
	BRIO	**27.46**	**40.64**	**50.50**	67.12	**27.95**	**37.27**	**43.68**	53.89
2	BART	4.81	12.76	22.05	42	5.41	11.01	16.89	30.02
	PEG	5.38	14.81	25.48	**48**	5.98	12.39	18.88	**33.79**
	BRIO	**6.18**	**15.53**	**25.44**	46.37	**7.75**	**13.68**	**19.01**	30.60
3	BART	0.95	4.68	11.28	29.40	1.93	5.11	9.32	20.97
	PEG	1.19	5.80	**13.93**	**35.03**	2.24	5.87	**10.57**	**23.99**
	BRIO	**1.47**	**6.50**	13.68	33.36	**2.86**	**6.35**	10.11	20.09

Table 7: Phrase overlap with the reference summary on n = 1, 2, and 3. For each n, we highlight minimum PO_{ref} and **maximum** PO_{ref} for a group within a dataset. A higher PO_{ref} suggests higher phrase overlap with the reference summary.

4 Conclusion

In this paper, we document our experiments on two widely used ATS datasets and three models trained on these datasets. We evaluate these on three dimensions of quality and demonstrate how the reported progress made in terms of the ROUGE metric is inconclusive. Our analysis shows that BART still shows competing behavior with current state-of-the-art models on various quality dimensions. We also highlight the need to carefully analyze the reference summaries in both datasets. Alternate evaluation metrics are required to account for different quality dimensions such as summarization complexity.

References

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting automatic evaluation of extractive summarization task: Can we do better than rouge? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560.

Chenxin An, Ming Zhong, Zhiyong Wu, Qin Zhu, Xuanjing Huang, and Xipeng Qiu. 2022. Colo: A contrastive learning based re-ranking framework for one-stage summarization. *arXiv preprint arXiv:2209.14569*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, and Pengfei Liu. 2020. Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5702–5711.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Freddy YY Choi. 2000. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309.

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

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.

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.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Wojciech Kryściński, 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.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen Mckeown. 2020. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive 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 2: Short Papers)*, pages 1065–1072.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

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.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the factual consistency of large language models through summarization. *arXiv preprint arXiv:2211.08412*.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3436–3449.

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.

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. 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 (EMNLP)*.

Topic 1: The security forces are reported to have used tear gas against stone-throwing protesters.

Topic 2: They also surrounded the hometown of Burhan Wani, 22, who was killed fighting Indian troops last year. Separately seven people are reported to have been killed in shelling across the Line of Control that divides Indian and Pakistani-administered Kashmir. Officials on the Pakistani side told Reuters that five people died in Indian shelling, while Indian officials say two people were killed by Pakistani fire.

Topic 3: There has been an armed revolt in the Muslim-majority region against rule by India since 1989, although violence has waned in recent years. The disputed region is claimed by both India and Pakistan in its entirety. India blames Pakistan for fuelling the unrest, a claim denied by Islamabad.

Topic 4: Burhan Wani is credited with reviving the image of militancy in Muslim-majority Indian-administered Kashmir, becoming a figurehead for young people. Saturday's violence started as people tried to walk to his home in Tral - where he died in a shootout with the army last July. His death led to a wave of protests during which dozens of people were killed.

Topic 5: The Indian authorities imposed heavy restrictions in the Kashmir valley for the anniversary, stopping internet access and sealing off Tral. There have also been reports of army personnel being injured in a militant attack overnight on Friday.

Figure 2: Example from the XSUM dataset. The article is segmented into five topics. Topic 1: Opening remark, Topic 2: Current situation on the incident, Topic 3: Background on India-Pakistan relationship, Topic 4: Background on the incident, Topic 5: Closing remark.

Article: Four years after becoming the youngest first-class cricketer in county history, Yorkshire's Barney Gibson has retired from the sport. The Leeds-born wicketkeeper entered the record books in 2011 when he lined up against Durham University just 27 days after his 15th birthday. But that match proved to be his only appearance at senior level and he never again progressed from the second XI. Ben Gibson, pictured at the age of 15, has decided to retire from cricket just four years after his debut . The 19-year-old said it was a 'difficult decision' to retire from cricket at such a young age . In his last game for the second string he did not bat or keep wicket, instead sending down 3.3 overs for 29 runs. 'This was a difficult decision to make,' the 19-year-old said. 'I would like to thank the players and staff at Yorkshire for their support. I have been involved with the club since I was 11 and I feel that now is the right time for me to look at a career change. 'The support from my parents has been tremendous and I would like to thank Ralph Middlebrook at Pudsey Congs Cricket Club and England coach Paul Farbrace, who I had close working relationships with.' Yorkshire's director of cricket development Ian Dews, said: 'Everyone at the club wishes Barney well. It is very much his decision. We hope that the next chapter in his life is very successful.'

Reference Summary: Barney Gibson became the youngest first-class cricketer in 2011 . The Yorkshire wicketkeeper made his debut shortly at 15 . Gibson said it was a 'difficult decision' to retire from the game .

Key-phrases: retired from the sport, cricketer in county history, youngest first-class cricketer, first-class cricketer in county, Yorkshire Barney Gibson, Barney Gibson has retired, Pudsey Congs Cricket Club, Durham University, Leeds-born wicketkeeper entered, England coach Paul Farbrace, Ralph Middlebrook at Pudsey, cricket development Ian Dews, lined up against Durham, cricket, wicketkeeper entered the record, entered the record books, difficult decision, Gibson

Figure 3: Example from the CNN-DM dataset. We highlight the key-phrase containing segments in the article. The key-phrases gives an overall idea about the important discussion points in the article.

ARTICLE: The Belgium international, 24, changed the game from the bench but fell awkwardly in injury time. His agent Patrick de Koster initially said De Bruyne would miss six weeks. But, after seeing a specialist, the £55m former Wolfsburg player said: “I’ll be out for around 10 weeks.” De Bruyne could miss up to 13 league and cup games, including the League Cup final with Liverpool on 28 February, both legs of the Champions League last-16 tie with Dynamo Kiev and the Manchester derby on 20 March. The Belgian is City’s second top goalscorer with 12 this season, four behind striker Sergio Aguero. De Koster added: “Kevin told me the only thing he can do is work hard and come back. Kevin is sad. His dream is to always be playing football.” De Bruyne scored one goal and set up another to help City to a 4-3 aggregate victory over the Toffees. Everton goalkeeper Joel Robles, who repeatedly tried to lift up De Bruyne as he lay injured, used social media to say sorry. "I would like to apologise to Kevin de Bruyne for my reaction to his injury," said the 25-year-old Spaniard. "In the heat of the moment I didn’t realise he was badly hurt. I wish him all the best and a speedy recovery.
REFERENCE SUMMARY: Manchester City midfielder Kevin de Bruyne says he will be out for about 10 weeks after injuring his right knee during Wednesday’s League Cup semi-final victory over Everton.

BART: Manchester City midfielder Kevin de Bruyne will be out for at least 10 weeks after injuring his ankle in Tuesday’s Champions League win over Everton.

PEGASUS: Manchester City midfielder Kevin de Bruyne will be out for up to 10 weeks with the ankle injury he suffered in Tuesday’s Capital One Cup win over Everton.

BRIO: Manchester City midfielder Kevin de Bruyne will be out for around 10 weeks after fracturing a bone in his right foot in the Capital One Cup win over Everton.

Figure 4: Example from the XSUM dataset. We underline the identified entities and highlight the entities with red that are missing from the source article.

	XSUM				CNN-DM			
	BART	PEGASUS	BRIO	Ref	BART	PEGASUS	BRIO	Ref
GPE	**24.34**	26.48	28.12	30.05	**0.58**	1.5	4.25	5.26
PERSON	**63.69**	63.7	66.28	67.32	**1.96**	9.41	24.4	8.44
ORG	**39.09**	40.85	45.66	45.97	**2.05**	8.5	18.39	10.92
DATE	**61.69**	66.06	67.76	76.71	**1.58**	2.88	5.78	18.71
CARDINAL	**42.17**	46.47	49.54	57.35	**0.31**	0.9	1.28	11.1
EVENT	**57.34**	60.21	62.16	58.8	**4.61**	11.45	23.53	17.49
LOC	**32.35**	38.05	44.4	46.74	**1.15**	2.69	15.56	10.36
ORDINAL	**34.57**	41.07	41.19	48.63	**0.63**	1.29	1.09	11.7
WORK_OF_ART	**38.71**	45.97	42.96	46.77	8.36	24.84	-	25.26
NORP	**26.64**	29.18	29.41	34.55	**0.9**	0.93	6.44	9.69
MONEY	**70.78**	72.61	77.07	86.21	**0.91**	1.57	2.79	16.84
PRODUCT	**25.42**	28.07	27.13	36.5	**0.24**	10.29	13.79	11.94
PERCENT	74.74	**67.44**	73.33	84.17	32.4	**25.0**	36.07	73.66
TIME	51.59	**50.0**	56.93	84.3	**3.08**	3.47	8.68	28.29
FAC	**54.79**	60.96	58.89	61.98	**2.97**	20.53	38.46	12.62
QUANTITY	**52.0**	69.23	77.42	94.38	1.38	**0.92**	3.11	20.18
LANGUAGE	12.5	-	12.5	44.44	-	-	27.78	10.1
LAW	66.67	**60.0**	69.05	70.83	**5.43**	34.07	20.0	35.48

Table 8: Class-wise EH distribution. We highlight maximum EH and **minimum EH** for an entity class within a dataset.