Neural Text Summarization: A Critical Evaluation

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann,
Caiming Xiong, Richard Socher
Salesforce Research
{kryscinski,nkeskar,bmccann,cxiong,rsocher}@salesforce.com

Abstract

Text summarization aims at compressing long documents into a shorter form that conveys the most important parts of the original document. Despite increased interest in the community and notable research effort, progress on benchmark datasets has stagnated. We critically evaluate key ingredients of the current research setup: datasets, evaluation metrics, and models, and highlight three primary shortcomings: 1) automatically collected datasets leave the task underconstrained and may contain noise detrimental to training and evaluation, 2) current evaluation protocol is weakly correlated with human judgment and does not account for important characteristics such as factual correctness, 3) models overfit to layout biases of current datasets and offer limited diversity in their outputs.

1 Introduction

Text summarization aims at compressing long textual documents into a short, human readable form that contains the most important information from the source. Two strategies of generating summaries are *extractive* (Dorr et al., 2003; Nallapati et al., 2017), where salient fragments of the source document are identified and directly copied into the summary, and *abstractive* (Rush et al., 2015; See et al., 2017), where the salient parts are detected and paraphrased to form the final output.

The number of summarization models introduced every year has been increasing rapidly. Advancements in neural network architectures (Sutskever et al., 2014; Bahdanau et al., 2015; Vinyals et al., 2015; Vaswani et al., 2017) and the availability of large scale data (Sandhaus, 2008; Nallapati et al., 2016a; Grusky et al., 2018) enabled the transition from systems based on expert knowledge and heuristics to data-driven approaches powered by end-to-end deep neural

models. Current approaches to text summarization utilize advanced attention and copying mechanisms (See et al., 2017; Tan et al., 2017; Cohan et al., 2018), multi-task and multi-reward training techniques (Guo et al., 2018; Pasunuru and Bansal, 2018; Kryściński et al., 2018), reinforcement learning strategies (Paulus et al., 2017; Narayan et al., 2018b; Dong et al., 2018; Wu and Hu, 2018), and hybrid extractive-abstractive models (Liu et al., 2018; Hsu et al., 2018; Gehrmann et al., 2018; Chen and Bansal, 2018). Many of the introduced models are trained on the CNN/DailyMail (Nallapati et al., 2016a) news corpus, a popular benchmark for the field, and are evaluated based on n-gram overlap between the generated and target summaries with the ROUGE package (Lin, 2004).

Despite substantial research effort, the progress on these benchmarks has stagnated. State-of-the-art models only slightly outperform the Lead-3 baseline, which generates summaries by extracting the first three sentences of the source document. We argue that this stagnation can be partially attributed to the current research setup, which involves uncurated, automatically collected datasets and non-informative evaluations protocols. We critically evaluate our hypothesis, and support our claims by analyzing three key components of the experimental setting: datasets, evaluation metrics, and model outputs. Our motivation is to shift the focus of the research community into developing a more robust research setup for text summarization.

2 Related Work

2.1 Datasets

To accommodate the requirements of modern data-driven approaches, several large-scale datasets have been proposed. The majority of

available corpora come from the news domain. Gigaword (Graff and Cieri, 2003) is a set of articles and corresponding titles that was originally used for headline generation (Takase et al., 2016), but it has also been adapted to single-sentence summarization (Rush et al., 2015; Chopra et al., 2016). NYT (Sandhaus, 2008) is a collection of articles from the New York Times magazine with abstracts written by library scientists. It has been primarily used for extractive summarization (Hong and Nenkova, 2014; Li et al., 2016) and phrase-importance prediction (Yang and Nenkova, 2014; Nye and Nenkova, 2015). The CNN/DailyMail (Nallapati et al., 2016a) dataset consists of articles with summaries composed of highlights from the article written by the authors themselves. It is commonly used for both *abstractive* (See et al., 2017; Paulus et al., 2017; Kryściński et al., 2018) and *extractive* (Dong et al., 2018; Wu and Hu, 2018; Zhou et al., 2018) neural summarization. The collection was originally introduced as a Cloze-style QA dataset by Hermann et al. (2015). XSum (Narayan et al., 2018a) is a collection of articles associated with one, single-sentence summary targeted at abstractive models. Newsroom (Grusky et al., 2018) is a diverse collection of articles sourced from 38 major online news outlets. This dataset was released together with a leaderboard and held-out testing split.

Outside of the news domain, several datasets were collected from open discussion boards and other portals offering structure information. Reddit TIFU (Kim et al., 2018) is a collection of posts scraped from Reddit where users post their daily stories and each post is required to contain a Too Long; Didn't Read (TL;DR) summary. Wiki-How (Koupaee and Wang, 2018) is a collection of articles from the WikiHow knowledge base, where each article contains instructions for performing procedural, multi-step tasks covering various areas, including: arts, finance, travel, and health.

2.2 Evaluation Metrics

Manual and semi-automatic (Nenkova and Passonneau, 2004; Passonneau et al., 2013) evaluation of large-scale summarization models is costly and cumbersome. Much effort has been made to develop automatic metrics that would allow for fast and cheap evaluation of models.

The ROUGE package (Lin, 2004) offers a set of automatic metrics based on the lexical over-lap between candidate and reference summaries. Overlap can be computed between consecutive (n-grams) and non-consecutive (skip-grams) subsequences of tokens. ROUGE scores are based on exact token matches, meaning that computing overlap between synonymous phrases is not supported.

Many approaches have extended ROUGE with support for synonyms and paraphrasing. ParaEval (Zhou et al., 2006) uses a three-step comparison strategy, where the first two steps perform optimal and greedy paraphrase matching based on paraphrase tables before reverting to exact token overlap. ROUGE-WE (Ng and Abrecht, 2015) replaces exact lexical matches with a soft semantic similarity measure approximated with the cosine distances between distributed representations of tokens. ROUGE 2.0 (Ganesan, 2018) leverages synonym dictionaries, such as WordNet, and considers all synonyms of matched words when computing token overlap. ROUGE-G (ShafieiBavani et al., 2018) combines lexical and semantic matching by applying graph analysis algorithms to the WordNet semantic network. Despite being a step in the direction of a more comprehensive evaluation protocol, none of these metrics gained sufficient traction in the research community, leaving ROUGE as the default automatic evaluation toolkit for text summarization.

2.3 Models

Existing summarization models fall into three categories: *abstractive*, *extractive*, and *hybrid*.

Extractive models select spans of text from the input and copy them directly into the summary. Non-neural approaches (Neto et al., 2002; Dorr et al., 2003; Filippova and Altun, 2013; Colmenares et al., 2015) utilized domain expertise to develop heuristics for summary content selection, whereas more recent, neural techniques allow for end-to-end training. In the most common case, models are trained as word- or sentence-level classifiers that predict whether a fragment should be included in the summary (Nallapati et al., 2016b, 2017; Narayan et al., 2017; Liu et al., 2019; Xu and Durrett, 2019). Other approaches apply reinforcement learning training strategies to directly optimize the model on task-specific, non-differentiable reward functions (Narayan et al., 2018b; Dong et al., 2018; Wu and Hu, 2018) .

Abstractive models paraphrase the source doc-

uments and create summaries with novel phrases not present in the source document. A common approach in *abstractive* summarization is to use attention and copying mechanisms (See et al., 2017; Tan et al., 2017; Cohan et al., 2018). Other approaches include using multi-task and multi-reward training (Paulus et al., 2017; Jiang and Bansal, 2018; Guo et al., 2018; Pasunuru and Bansal, 2018; Kryściński et al., 2018), and unsupervised training strategies (Chu and Liu, 2018; Schumann, 2018).

Hybrid models (Hsu et al., 2018; Liu et al., 2018; Gehrmann et al., 2018; Chen and Bansal, 2018) include both *extractive* and *abstractive* modules and allow to separate the summarization process into two phases – content selection and paraphrasing.

For the sake of brevity we do not describe details of different models, we refer interested readers to the original papers.

2.4 Analysis and Critique

Most summarization research revolves around new architectures and training strategies that improve the state of the art on benchmark problems. However, it is also important to analyze and question the current methods and research settings.

Zhang et al. (2018) conducted a quantitative study of the level of abstraction in *abstractive* summarization models and showed that word-level, copy-only *extractive* models achieve comparable results to fully *abstractive* models in the measured dimension. Kedzie et al. (2018) offered a thorough analysis of how neural models perform content selection across different data domains, and exposed data biases that dominate the learning signal in the news domain and architectural limitations of current approaches in learning robust sentence-level representations. Liu and Liu (2010) examine the correlation between ROUGE scores and human judgments when evaluating meeting summarization data and show that the correlation strength is low, but can be improved by leveraging unique meeting characteristics, such as available speaker information. Owczarzak et al. (2012) inspect how inconsistencies in human annotator judgments affect the ranking of summaries and correlations with automatic evaluation metrics. The results showed that system-level rankings, considering all summaries, were stable despite inconsistencies in judgments, however, summary-level rankings and automatic metric correlations benefit from improving annotator consistency. Graham (2015) compare the fitness of the BLEU metric (Papineni et al., 2002) and a number of different ROUGE variants for evaluating summarization outputs. The study reveals superior variants of ROUGE that are different from the commonly used recommendations and shows that the BLEU metric achieves strong correlations with human assessments of generated summaries. Schulman et al. (2015) study the problems related to using ROUGE as an evaluation metric with respect to finding optimal solutions and provide proof of NP-hardness of global optimization with respect to ROUGE.

Similar lines of research, where the authors put under scrutiny existing methodologies, datasets, or models were conducted by Callison-Burch et al. (2006, 2007); Tan et al. (2015); Post (2018) in machine translation, (Gkatzia and Mahamood, 2015; Reiter and Belz, 2009; Reiter, 2018) in natural language generation, Lee et al. (2016); Chen et al. (2016); Kaushik and Lipton (2018) in reading comprehension, Gururangan et al. (2018); Poliak et al. (2018); Glockner et al. (2018) in natural language inference, Goyal et al. (2017) in visual question answering, and Xian et al. (2017) in zero-shot image classification. Comments on the general state of scholarship in the field of machine learning were presented by Sculley et al. (2018); Lipton and Steinhardt (2019) and references therein.

3 Datasets

3.1 Underconstrained task

The task of summarization is to compress long documents by identifying and extracting the most important information from the source documents. However, assessing the importance of information is a difficult task in itself, that highly depends on the expectations and prior knowledge of the target reader.

We show that the current setting in which models are simply given a document with one associated reference summary and no additional information, leaves the task of summarization underconstrained and thus too ambiguous to be solved by end-to-end models.

To quantify this effect, we conducted a human study which measured the agreement between different annotators in selecting important sentences

Article	
The glowing blue letters that once lit the Bronx from above Yankee stadium failed to find a buyer at an auction at Sotheby's on Wednesday. While the 13 letters were expected to bring in anywhere from $300,000 to $600,000, the only person who raised a paddle - for $260,000 - was a Sotheby's employee trying to jump start the bidding. The current owner of the signage is Yankee hall-of-famer Reggie Jackson, who purchased the 10-feet-tall letters for an undisclosed amount after the stadium saw its final game in 2008. No love: 13 letters that hung over Yankee stadium were estimated to bring in anywhere from $300,000 to $600,000, but received no bids at a Sotheby's auction Wednesday. The 68-year-old Yankee said he wanted 'a new generation to own and enjoy this icon of the Yankees and of New York City.', The letters had beamed from atop Yankee stadium near grand concourse in the Bronx since 1976, the year before Jackson joined the team. (...)	

Summary Questions	
When was the auction at Sotheby's?	
Who is the owner of the signage?	
When had the letters been installed on the stadium?	

Constrained Summary A	Unconstrained Summary A
Glowing letters that had been hanging above the Yankee stadium from 1976 to 2008 were placed for auction at Sotheby's on Wednesday, but were not sold, The current owner of the sign is Reggie Jackson, a Yankee hall-of-famer.	There was not a single buyer at the auction at Sotheby's on Wednesday for the glowing blue letters that once lit the Bronx's Yankee Stadium. Not a single non-employee raised their paddle to bid. Jackson, the owner of the letters, was surprised by the lack of results. The venue is also auctioning off other items like Mets memorabilia.

Constrained Summary B	Unconstrained Summary B
An auction for the lights from Yankee Stadium failed to produce any bids on Wednesday at Sotheby's. The lights, currently owned by former Yankees player Reggie Jackson, lit the stadium from 1976 until 2008.	The once iconic and attractive pack of 13 letters that was placed at the Yankee stadium in 1976 and later removed in 2008 was unexpectedly not favorably considered at the Sotheby's auction when the 68 year old owner of the letters attempted to transfer its ownership to a member the younger populace. Thus, when the minimum estimate of $300,000 was not met, a further attempt was made by a former player of the Yankees to personally visit the new owner as an

Table 1: Example summaries collected from human annotators in the *constrained* (left) and *unconstrained* (right) task. In the *unconstrained* setting, annotators were given a news article and asked to write a summary covering the parts they considered most important. In the *constrained* setting, annotators were given a news article with three associated questions and asked to write a summary that contained the answers to the given questions.

from a fragment of text. We asked workers to write summaries of news articles and highlight sentences from the source documents that they based their summaries on. The experiment was conducted in two settings: *unconstrained*, where the annotators were instructed to summarize the content that they considered most important, and *constrained*, where annotators were instructed to write summaries that would contain answers to three questions associated with each article. This is similar to the construction of the TAC 2008 Opinion Summarization Task [1]. The questions associated with each article where collected from human workers through a separate assignment. Experiments were conducted on 100 randomly sampled articles, further details of the human study can be found in Appendix A.1.

Table 2 shows the average number of sentences, per-article, that annotators agreed were important.

The rows show how the average changes with the human vote threshold needed to reach consensus about the importance of any sentence. For example, if we require that three or more human votes are necessary to consider a sentence important, annotators agreed on average on the importance of 0.627 and 1.392 sentences per article in the *unconstrained* and *constrained* settings respectively. The average length (in sentences) of sampled articles was 16.59, with a standard deviation of 5.39. The study demonstrates the difficulty and ambiguity of content selection in text summarization.

We also conducted a qualitative study of summaries written by annotators. Examples comparing summaries written in the *constrained* and *unconstrained* setting are shown in Table 1. We noticed that in both cases the annotators correctly identified the main topic and important fragments of the source article. However, *constrained* summaries were more succinct and targeted, without sacrificing the natural flow of sentences. *Un-*

Human vote threshold	Sent. per article considered important	
	Unconstrained	*Constrained*
$= 5$	0.028	0.251
≥ 4	0.213	0.712
≥ 3	0.627	1.392
≥ 2	1.695	2.404
≥ 1	5.413	4.524

Table 2: Average number of sentences, per-article, which annotators agreed were important. The human vote threshold investigates how the average agreement changes with the threshold of human votes required to consider any sentence important. Rows $= 5$ and ≥ 1 correspond to the set intersection and union of selected sentences accordingly.

constrained writers tended to write more verbose summaries that did not add information. The study also highlights the *abstractive* nature of human written summaries in that similar content can be described in unique ways.

3.2 Layout bias in news data
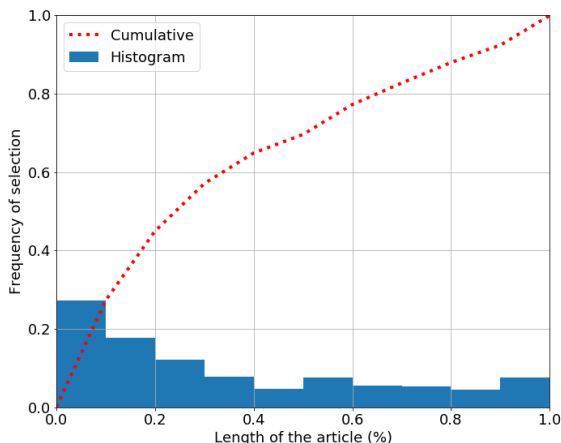
Figure 1: The distribution of important sentences over the length of the article according to human annotators (blue) and its cumulative distribution (red).

News articles adhere to a writing structure known in journalism as the "Inverted Pyramid" (PurdueOWL, 2019). In this form, initial paragraphs contain the most newsworthy information, which is followed by details and background information.

To quantify how strongly articles in the CNN/DM corpus follow this pattern we conducted a human study that measured the importance of different sections of the article. Annotators read news articles and selected sentences they found

most important. Experiments were conducted on 100 randomly sampled articles, further details of the human study are described in Appendix A.3. Figure 1 presents how annotator selections were distributed over the length of the article. The distribution is skewed towards the first quarter of the length of articles. The cumulative plot shows that nearly 60% of the important information was present in the first third of the article, and approximately 25% and 15% of selections pointing to the second and last third, respectively.

It has become standard practice to exploit such biases during training to increase performance of models (See et al., 2017; Paulus et al., 2017; Kryściński et al., 2018; Gehrmann et al., 2018; Jiang and Bansal, 2018; Pasunuru and Bansal, 2018), but the importance of these heuristics has been accepted without being quantified. These same heuristics would not apply to books or legal documents, which lack the Inverted Pyramid layout so common in the news domain, so it is important that these heuristics be part of ablation studies rather than accepted as default pre-processing step.

3.3 Noise in scraped datasets

Given the data requirements of deep neural networks and the vast amounts of diverse resources available online, automatically scraping web content is a convenient way of collecting data for new corpora. However, adapting scraped content to the needs of end-to-end models is problematic. Given that manual inspection of data is infeasible and human annotators are expensive, data curation is usually limited to removing any markup structure and applying simple heuristics to discard obviously flawed examples. This, in turn, makes the quality of the datasets heavily dependent on how well the scraped content adheres to the assumptions made by the authors about its underlying structure.

This issue suggests that available summarization datasets would be filled with noisy examples. Manual inspection of the data, particularly the reference summaries, revealed easily detectable, consistent patterns of flawed examples Many such examples can be isolated using simple regular expressions and heuristics, which allows approximation of how widespread these flaws are in the dataset.

We investigated this issue in two large summarization corpora scraped from the internet:

CNN/DM - Links to other articles
Michael Carrick has helped Manchester United win their last six games. Carrick should be selected alongside Gary Cahill for England. Carrick has been overlooked too many times by his country. READ : Carrick and Man United team-mates enjoy second Christmas party.

Newsroom - Links to news sources
Get Washington DC, Virginia, Maryland and national news. Get the latest/breaking news, featuring national security, science and courts. Read news headlines from the nation and from The Washington Post. Visit www.washingtonpost.com/nation today.

Table 3: Examples of noisy reference summaries found in the CNN/DM and Newsroom datasets.

Article
Quick-thinking: Brady Olson, a teacher at North Thurston High, took down a gunman on Monday. A Washington High School teacher is being hailed a hero for tackling a 16-year-old student to the ground after he opened fire on Monday morning (...)

Summary - Factually incorrect
Brady Olson, a Washington High School teacher at North Thurston High, opened fire on Monday morning. No one was injured after the boy shot twice toward the ceiling in the school commons before classes began at North Thurston High School in Lacey (...)

Table 4: Example of a factually incorrect summary generated by an abstractive model. Top: ground-truth article. Bottom: summary generated by model.

CNN/DM (Nallapati et al., 2016a) and the Newsroom (Grusky et al., 2018). The problem of noisy data affects 0.47%, 5.92%, and 4.19% of the training, validation, and test split of the CNN/DM dataset, and 3.21%, 3.22%, and 3.17% of the respective splits of the Newsroom dataset. Examples of noisy summaries are shown in Table 3. Flawed examples contained links to other articles and news sources, placeholder texts, unparsed HTML code, and non-informative passages in the reference summaries.

4 Evaluation Metrics

4.1 Weak correlation with human judgment

The effectiveness of ROUGE was previously evaluated (Lin, 2004; Graham, 2015) through statistical correlations with human judgment on the DUC datasets (Over and Yen, 2001, 2002, 2003). However, their setting was substantially different from the current environment in which summarization models are developed and evaluated.

To investigate the robustness of ROUGE in the setting in which it is currently used, we evaluate how its scores correlate with the judgment of an average English-speaker using examples from the CNN/DM dataset. Following the human evaluation protocol from Gehrmann et al. (2018), we asked annotators to rate summaries across four dimensions: *relevance* (selection of important content from the source), *consistency* (factual alignment between the summary and the source), *fluency* (quality of individual sentences), and *coherence* (collective quality of all sentences). Each summary was rated by 5 distinct judges with the final score obtained by averaging the individual scores. Experiments were conducted on 100 randomly sampled articles with the outputs of 13 summarization systems provided by the original authors. Correlations were computed between all pairs of Human-, ROUGE-scores, for all systems. Additional summaries were collected from annotators to inspect the effect of using multiple ground-truth labels on the correlation with automatic metrics. Further details of the human study can be found in Appendix A.2.

Results are shown in Table 5. The left section of the table presents Pearson's correlation coefficients and the right section presents Kendall rank correlation coefficients. In terms of Pearsons's coefficients, the study showed minimal correlation with any of the annotated dimensions for both *abstractive* and *extractive* models together and for *abstractive* models individually. Weak correlation was discovered for *extractive* models primarily with the *fluency* and *coherence* dimensions.

We hypothesized that the noise contained in the fine-grained scores generated by both human annotators and ROUGE might have affected the correlation scores. We evaluated the relation on a higher level of granularity by means of correlation between rankings of models that were obtained from the fine-grained scores. The study showed weak correlation with all measured dimensions, when evaluated for both *abstractive* and *extractive* models together and for *abstractive* models individually. Moderate correlation was found for *extractive* models across all dimensions. A surprising result was that correlations grew weaker with the increase of ground truth references.

Our results align with the observations from Liu and Liu (2010) who also evaluated ROUGE outside of its original setting. The study highlights the limited utility in measuring progress of the field

	Pearson correlation									Kendall rank correlation								
	1 Reference			5 References			10 References			1 Reference			5 References			10 References		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
All Models																		
Relevance	0.07	0.03	0.06	0.03	0.02	0.02	0.05	0.03	0.04	0.29	0.30	0.29	0.28	0.29	0.27	0.28	0.28	0.27
Consistency	0.08	0.03	0.07	0.02	0.01	0.01	0.03	0.01	0.02	0.27	0.27	0.28	0.28	0.28	0.27	0.28	0.30	0.28
Fluency	0.08	0.06	0.08	0.05	0.03	0.04	0.05	0.04	0.05	0.28	0.27	0.29	0.26	0.28	0.27	0.26	0.26	0.25
Coherence	0.06	0.05	0.07	0.05	0.04	0.05	0.04	0.03	0.04	0.29	0.31	0.29	0.27	0.27	0.27	0.28	0.28	0.27
Abstractive Models																		
Relevance	0.04	0.01	0.05	0.01	0.00	0.00	0.04	0.02	0.03	0.31	0.28	0.29	0.25	0.28	0.27	0.28	0.26	0.24
Consistency	0.07	0.01	0.06	0.00	−0.02	−0.01	0.03	0.01	0.03	0.30	0.30	0.29	0.26	0.25	0.25	0.24	0.27	0.24
Fluency	0.06	0.04	0.07	0.03	0.01	0.02	0.05	0.04	0.04	0.28	0.28	0.28	0.24	0.25	0.24	0.27	0.26	0.24
Coherence	0.04	0.02	0.04	0.02	0.01	0.02	0.03	0.02	0.03	0.30	0.32	0.31	0.25	0.27	0.26	0.28	0.29	0.26
Extractive Models																		
Relevance	0.14	0.09	0.13	0.09	0.05	0.07	0.06	0.03	0.04	0.45	0.44	0.49	0.38	0.38	0.38	0.44	0.43	0.42
Consistency	0.10	0.09	0.11	0.07	0.07	0.07	0.00	−0.03	−0.02	0.48	0.49	0.49	0.38	0.39	0.38	0.42	0.41	0.42
Fluency	0.13	0.14	0.13	0.10	0.10	0.08	0.06	0.03	0.04	0.47	0.47	0.48	0.35	0.37	0.35	0.39	0.34	0.37
Coherence	0.15	0.17	0.15	0.13	0.13	0.13	0.08	0.05	0.06	0.42	0.43	0.46	0.40	0.39	0.39	0.38	0.34	0.36

Table 5: Correlations between human annotators and ROUGE scores along different dimensions and multiple reference set sizes. Left: Pearson's correlation coefficients. Right: Kendall's rank correlation coefficients.

solely by means of ROUGE scores.

4.2 Insufficient evaluation protocol

The goal of text summarization is to automatically generate succinct, fluent, relevant, and factually consistent summaries. The current evaluation protocol depends primarily on the exact lexical overlap between reference and candidate summaries measured by ROUGE. In certain cases, ROUGE scores are complemented with human studies where annotators rate the *relevance* and *fluency* of generated summaries. Neither of the methods explicitly examines the factual consistency of summaries, leaving this important dimension unchecked.

To evaluate the factual consistency of existing models, we manually inspected randomly sampled articles with summaries coming from randomly chosen, *abstractive* models. We focused exclusively on factual incorrectness and ignored any other issues, such as low fluency. Out of 200 article-summary pairs that were reviewed manually, we found that 60 (30%) contained consistency issues. Table 4 shows examples of discovered inconsistencies. Some of the discovered inconsistencies, despite being factually incorrect, could be rationalized by humans. However, in many cases, the errors were substantial and could have severe repercussions if presented as-is to target readers.

5 Models

5.1 Layout bias in news data

We revisit the problem of layout bias in news data from the perspective of models. Kedzie et al.

(2018) showed that in the case of news articles, the layout bias dominates the learning signal for neural models. In this section, we approximate the degree with which generated summaries rely on the leading sentences of news articles.

We computed ROUGE scores for collected models in two settings: first using the CNN/DM reference summaries as the ground-truth, and second where the leading three sentences of the source article were used as the ground-truth, i.e. the Lead-3 baseline. We present the results in Table 6.

For all examined models we noticed a substantial increase of overlap across all ROUGE variants. Results suggest that performance of current models is strongly affected by the layout bias of news corpora. Lead-3 is a strong baseline that exploits the described layout bias. However, there is still a large gap between its performance and an upper bound for extractive models (extractive oracle).

5.2 Diversity of model outputs

Models analyzed in this paper are considerably different from each other in terms of architectures, training strategies, and underlying approaches. We inspected how the diversity in approaches translates into the diversity of model outputs.

We computed ROUGE-1 and ROUGE-4 scores between pairs of model outputs to compare them by means of token and phrase overlap. Results are visualized in Figure 2, where the values above and below the diagonal are ROUGE-1 and -4 scores accordingly, and model names (M-) follow the order from Table 6.

	Target Reference					Lead-3 Reference				
	R-1	**R-2**	**R-3**	**R-4**	**R-L**	**R-1**	**R-2**	**R-3**	**R-4**	**R-L**
Extractive Oracle (Grusky et al., 2018)	93.36	83.19	-	-	93.36	-	-	-	-	-
Lead-3 Baseline	40.24	17.53	9.94	6.50	36.49	-	-	-	-	-
Abstractive Models										
Model Hsu et al. (2018)	40.68	17.97	10.43	6.97	37.13	69.66	62.60	60.33	58.72	68.42
Model Gehrmann et al. (2018)	41.53	18.77	10.68	6.98	38.39	52.25	39.03	33.40	29.61	50.21
Model Jiang and Bansal (2018)	40.05	17.66	10.34	6.99	36.73	62.32	52.93	49.95	47.98	60.72
Model Chen and Bansal (2018)	40.88	17.81	9.79	6.19	38.54	55.87	41.30	34.69	29.88	53.83
Model See et al. (2017)	39.53	17.29	10.05	6.75	36.39	58.15	47.60	44.11	41.82	56.34
Model Kryściński et al. (2018)	40.23	17.30	9.33	5.70	37.76	57.22	42.30	35.26	29.95	55.13
Model Li et al. (2018)	40.78	17.70	9.76	6.19	38.34	56.45	42.36	35.97	31.39	54.51
Model Pasunuru and Bansal (2018)	40.44	18.03	10.56	7.12	37.02	62.81	53.57	50.25	47.99	61.27
Model Zhang et al. (2018)	39.75	17.32	10.11	6.83	36.54	58.82	47.55	44.07	41.84	56.83
Model Guo et al. (2018)	39.81	17.64	10.40	7.08	36.49	56.42	45.88	42.39	40.11	54.59
Extractive Models										
Model Dong et al. (2018)	41.41	18.69	10.87	7.22	37.61	73.10	66.98	65.49	64.66	72.05
Model Wu and Hu (2018)	41.25	18.87	11.05	7.38	37.75	78.68	74.74	73.74	73.12	78.08
Model Zhou et al. (2018)	41.59	19.00	11.13	7.45	38.08	69.32	61.00	58.51	56.98	67.85

Table 6: ROUGE (R-) scores computed for different models on the test set of the CNN/DM dataset. Left: Scores computed with the original reference summaries. Right: Scores computed with Lead-3 used as the reference.

We notice that the ROUGE-1 scores vary considerably less than ROUGE-4 scores. This suggests that the models share a large part of the vocabulary on the token level, but differ on how they organize the tokens into longer phrases.

Comparing results with the n-gram overlap between models and reference summaries (Table 6) shows a substantially higher overlap between any model pair than between the models and reference summaries. This might imply that the training data contains easy to pick up patterns that all models overfit to, or that the information in the training signal is too weak to connect the content of the source articles with the reference summaries.

6 Conclusions

This critique has highlighted the weak points of the current research setup in text summarization. We showed that text summarization datasets require additional constraints to have well-formed summaries, current state-of-the-art methods learn to rely too heavily on layout bias associated with the particular domain of the text being summarized, and the current evaluation protocol reflects human judgments only weakly while also failing to evaluate critical features (e.g. factual correctness) of text summarization.

We hope that this critique provides the summarization community with practical insights for future research directions that include the construction of datasets, models less fit to a particular do-
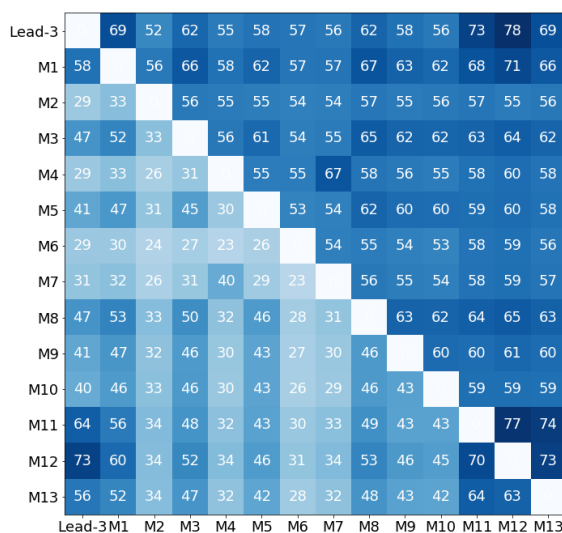

Figure 2: Pairwise similarities between model outputs computed using ROUGE. Above diagonal: Unigram overlap (ROUGE-1). Below diagonal: 4-gram overlap (ROUGE-4). Model order (M-) follows Table 6.

main bias, and evaluation that goes beyond current metrics to capture the most important features of summarization.

7 Acknowledgements

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Chris Callison-Burch, Cameron S. Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *WMT@ACL*, pages 136–158. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *EACL*. The Association for Computer Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *ACL (1)*. The Association for Computer Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *ACL (1)*, pages 675–686. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*.

Eric Chu and Peter J. Liu. 2018. Unsupervised neural multi-document abstractive summarization. *CoRR*, abs/1810.05739.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*.

Carlos A Colmenares, Marina Litvak, Amin Mantrach, and Fabrizio Silvestri. 2015. Heads: Headline generation as sequence prediction using an abstract feature-rich space. In *HLT-NAACL*, pages 133–142.

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, Brussels, Belgium, October 31 - November 4, 2018*.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *HLT-NAACL*.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of EMNLP*, pages 1481–1491. Citeseer.

Kavita Ganesan. 2018. ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks. *CoRR*, abs/1803.01937.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In *EMNLP*, pages 4098–4109. Association for Computational Linguistics.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *ENLG 2015 - Proceedings of the 15th European Workshop on Natural Language Generation, 10-11 September 2015, University of Brighton, Brighton, UK*, pages 57–60.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *ACL (2)*, pages 650–655. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334. IEEE Computer Society.

David Graff and C Cieri. 2003. English gigaword, linguistic data consortium.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *EMNLP*, pages 128–137. The 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*, pages 107–112. 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 *NIPS*.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.

Yichen Jiang and Mohit Bansal. 2018. Closed-book training to improve summarization encoder memory. In *EMNLP*, pages 4067–4077. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *EMNLP*, pages 5010–5015. Association for Computational Linguistics.

Chris Kedzie, Kathleen R. McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *EMNLP*, pages 1818–1828. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2018. Abstractive summarization of reddit posts with multi-level memory networks. *CoRR*, abs/1811.00783.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *EMNLP*, pages 1808–1817. Association for Computational Linguistics.

Moontae Lee, Xiaodong He, Wen-tau Yih, Jianfeng Gao, Li Deng, and Paul Smolensky. 2016. Reasoning in vector space: An exploratory study of question answering. In *ICLR*.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018. Improving neural abstractive document summarization with structural regularization. In *EMNLP*, pages 4078–4087. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship. *ACM Queue*, 17(1):80.

Feifan Liu and Yang Liu. 2010. Exploring correlation between ROUGE and human evaluation on meeting summaries. *IEEE Trans. Audio, Speech & Language Processing*, 18(1):187–196.

Jingyun Liu, Jackie Chi Kit Cheung, and Annie Louis. 2019. What comes next? extractive summarization by next-sentence prediction. *CoRR*, abs/1901.03859.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Ramesh Nallapati, Bowen Zhou, Çağlar Gülçehre, Bing Xiang, et al. 2016a. Abstractive text summarization using sequence-to-sequence rnns and beyond. *Proceedings of SIGNLL Conference on Computational Natural Language Learning*.

Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016b. Classify or select: Neural architectures for extractive document summarization. *CoRR*, abs/1611.04244.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. 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*, Brussels, Belgium.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL-HLT*, pages 1747–1759. Association for Computational Linguistics.

Shashi Narayan, Nikos Papasarantopoulos, Mirella Lapata, and Shay B. Cohen. 2017. Neural extractive summarization with side information. *CoRR*, abs/1704.04530.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*.

Joel Larocca Neto, Alex A Freitas, and Celso AA Kaestner. 2002. Automatic text summarization using a machine learning approach. In *Brazilian Symposium on Artificial Intelligence*, pages 205–215. Springer.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. *CoRR*, abs/1508.06034.

Benjamin Nye and Ani Nenkova. 2015. Identification and characterization of newsworthy verbs in world news. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015.*

Paul Over and James Yen. 2001. An introduction to duc-2001: Intrinsic evaluation of generic news text summarization systems.

Paul Over and James Yen. 2002. An introduction to duc-2002: Intrinsic evaluation of generic news text summarization systems.

Paul Over and James Yen. 2003. An introduction to duc-2003: Intrinsic evaluation of generic news text summarization systems.

Karolina Owczarzak, Peter A. Rankel, Hoa Trang Dang, and John M. Conroy. 2012. Assessing the effect of inconsistent assessors on summarization evaluation. In *ACL (2)*, pages 359–362. The Association for Computer Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Rebecca J. Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated pyramid scoring of summaries using distributional semantics. In *ACL (2)*, pages 143–147. The Association for Computer Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. *CoRR*, abs/1804.06451.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. In *ICLR*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In **SEM@NAACL-HLT*, pages 180–191. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *WMT*, pages 186–191. Association for Computational Linguistics.

PurdueOWL. 2019. Journalism and journalistic writing: The inverted pyramid structure. Accessed: 2019-05-15.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3).

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of EMNLP*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. 2015. Gradient estimation using stochastic computation graphs. In *NIPS*.

Raphael Schumann. 2018. Unsupervised abstractive sentence summarization using length controlled variational autoencoder. *CoRR*, abs/1809.05233.

D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and Ali Rahimi. 2018. Winner's curse? on pace, progress, and empirical rigor. In *ICLR (Workshop)*. OpenReview.net.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond K. Wong, and Fang Chen. 2018. A graph-theoretic summary evaluation for rouge. In *EMNLP*, pages 762–767. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016.*

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers.*

Liling Tan, Jon Dehdari, and Josef van Genabith. 2015. An awkward disparity between BLEU / RIBES scores and human judgements in machine translation. In *WAT*, pages 74–81. Workshop on Asian Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *NIPS*.

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*.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, pages 3077–3086. IEEE Computer Society.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. *CoRR*, abs/1902.00863.

Yinfei Yang and Ani Nenkova. 2014. Detecting information-dense texts in multiple news domains. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*.

Fangfang Zhang, Jin-ge Yao, and Rui Yan. 2018. On the abstractiveness of neural document summarization. In *EMNLP*, pages 785–790. Association for Computational Linguistics.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard H. Hovy. 2006. Paraeval: Using paraphrases to evaluate summaries automatically. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.