Intrinsic Evaluation of Summarization Datasets

Rishi Bommasani
Department of Computer Science
Stanford University
nlprishi@cs.stanford.edu

Claire Cardie
Department of Computer Science
Cornell University
cardie@cs.cornell.edu

Abstract

High quality data forms the bedrock for building meaningful statistical models in NLP. Consequently, data quality must be evaluated either during dataset construction or *post hoc*. Almost all popular summarization datasets are drawn from natural sources and do not come with inherent quality assurance guarantees. In spite of this, data quality has gone largely unquestioned for many recent summarization datasets. We perform the first large-scale evaluation of summarization datasets by introducing 5 intrinsic metrics and applying them to 10 popular datasets. We find that data usage in recent summarization research is sometimes inconsistent with the underlying properties of the datasets employed. Further, we discover that our metrics can serve the additional purpose of being inexpensive heuristics for detecting generically low quality examples.

1 Introduction

Data understanding is fundamentally important in natural language processing (NLP); for data-driven learning-based methods (e.g. neural networks), the quality of the training data bounds the quality of models learned using it. Therefore, understanding this data is necessary in order to ensure that models learn to perform a given task correctly.

Understanding data is a multidimensional problem. One line of inquiry has demonstrated why prominent datasets are insufficiently challenging: many data examples can be solved by alternative heuristics that do not encode an approach that is faithful to the task (McCoy et al., 2019). From the perspective of datasets, several works have shown that standard datasets in areas such as visual question answering (Zhang et al., 2016; Kafle and Kanan, 2017), natural language inference (Gururangan et al., 2018; Poliak et al., 2018), and reading comprehension (Kaushik and Lipton, 2018) contain annotation artifacts that often give rise to these spurious correlations or reasoning shortcuts. Data understanding can also inform scientific and ethical decision-making (Bender and Friedman, 2018; Gebru et al., 2018; Mitchell et al., 2019) with recent work studying how social biases encoded in training data propagate to learned models (Zhao et al., 2019; Tan and Celis, 2019).

In this work, we extend these efforts towards the setting of summarization. We find this to be particularly timely since several summarization datasets have been released in recent years with little discussion of data quality. While prior work on evaluating NLP datasets has focused on their difficulty, transparency, or bias, we consider broadly the overall quality of the dataset — in our case, for the task of summarization. Our central insight is that desirable properties of a summary can be readily estimated by adapting and applying existing NLP methods. With this in mind, we present a multi-aspect large-scale study of summarization datasets that dissects summarization into 5 properties that are evaluated across 10 datasets spanning multiple summarization domains. Our analysis reveals that our metrics can serve as lightweight detectors of generically low quality examples. Most strikingly, we show that quantifiable aspects of summarization datasets are inconsistent with their use by the NLP community in several instances.

2 Motivation

Quality assurance for data. Nuanced understanding of data is requisite for drawing sound scientific conclusions. In particular, without evaluating for the quality and accuracy of data used to test models, it is impossible to be certain that progress is being made and that successive iterations of models truly make progress on the underlying task or linguistic phenomena of interest.

Within NLP, iconic datasets such as the Penn

Treebank (Marcus et al., 1993) have sustained subareas such as language modelling, part-of-speech tagging, and syntactic parsing for years due to the painstaking annotation efforts put into making these high-fidelity resources. And in the context of summarization, initial datasets, such as those produced during the Document Understanding Conference (DUC) and Text Analysis Conference (TAC) evaluations, implemented fine-grained verification of data quality.[1]

In part due to the emergence of data-hungry modelling techniques, the demands for larger datasets often render quality assurance procedures of this standard to be impractical and infeasible. Nonetheless, several recent natural language understanding datasets (Bowman et al., 2015; Rajpurkar et al., 2016; Suhr et al., 2017) institute explicit quality-control procedures in crowd-sourcing dataset construction (Zaidan and Callison-Burch, 2011; Yan et al., 2014; Callison-Burch et al., 2015), such as using additional annotators to validate annotations (c.f. Geva et al., 2019). In the sibling subfield of machine translation, which often shares similar modelling challenges and evaluation regimes as summarization due to the shared nature of being sequence-to-sequence natural language generation tasks, the annual WMT conference[2] consistently furnishes high quality data. In summary, ensuring data quality is both crucial and challenging. And in comparison with other subareas of NLP, we argue that summarization has lagged behind in rigorously ensuring the quality of widely-used datasets.

Relating data quality and model quality. The correctness and quality of data inherently bounds what can be learned from the data about the task of interest. From an information-theoretic perspective, this can be made fully formal as follows:[3]

$$\underbrace{I(S;M)}_{\text{learned model}} \leq \underbrace{I(S;T)}_{\text{training data}} + \underbrace{I(S;P)}_{\text{pretraining}} + \underbrace{I(S;A)}_{\text{inductive bias}}$$

Here, I denotes the mutual information, S denotes understanding of the underlying summarization task and M denotes a model learned using summarization training data T, additional pretraining data P, and the model's architecture A. For fully learning-based methods, especially those with weak/minimal inductive biases such as neural networks, $I(S;A)$ is approximately zero. While $I(S;P)$ may be greater than zero (e.g. language modelling pretraining provides statistical information that may facilitate a model to avoid *a priori* unlikely summaries), standard pretraining regimes such as large-scale language modelling over generic text corpora (Devlin et al., 2019; Raffel et al., 2019) are likely insufficient to meaningfully learn to summarize. Under these assumptions, the mutual information between S and M is critically upper-bounded in terms of $I(S;T)$. We hypothesize that the quality of the training dataset T is highly correlated with its mutual information with respect to the summarization task S, $I(S;T)$.

One size does not fit all. Spärck Jones (1999) famously argued that summarization systems should be understood conditional on the context in which they will be used. In recent years, the field has significantly departed from this perspective and primarily studied "general-purpose summarization" (Kryscinski et al., 2019), which she denounced as *ignis fatuus*. With our work, we adopt the perspective that for all datasets it is strictly preferable to have all properties quantified; it is the responsibility of practitioners building summarization systems to accurately weight different metrics based on their ultimate goals and use cases. As such, we refrain from providing prescriptive domain-agnostic or context-agnostic notions of summarization.

3 Metrics

In this work, we evaluate the quality of a dataset by aggregating scores for each example in the dataset. We conjecture that for many NLP tasks, estimating the quality of a particular data example is of similar complexity as correctly performing the task on the example.[4] Nevertheless, for summarization, our insight is that various aspects of a summarization example (a document-summary pair) can be reliably estimated by re-purposing existing NLP methods. We are guided by pioneering work (Luhn, 1958; Edmundson, 1969; Mani, 1999) that defined core properties of summarization systems and influential subsequent work (Radev et al., 2002; Nenkova, 2006; Nenkova and McKeown, 2012; Peyrard, 2019a) that refined and extended these properties. From

[1] DUC 2003 annotation guidelines: https://duc.nist.gov/duc2003/tasks.html and DUC 2002 quality assessment questions: https://duc.nist.gov/duc2003/quality.html

[2] http://www.statmt.org/wmt20/

[3] Proof deferred to Appendix D.

[4] Research in algorithms provides a natural parallel: many computationally hard optimization problems remain intractable when relaxed to their decision problem version. For example, the travelling salesman problem of finding the least costly Hamiltonian cycle remains NP-hard even if we just ask *"Does there exist a Hamiltonian cycle of cost $\leq L$?"*

this literature, we specifically study *compression*, *topic similarity*, *abstractivity*, *redundancy*, and *semantic coherence* as these properties are of recurring and sustained interest.[5]

For each abstract property, numerous concrete methods can be proposed to quantify it. In Appendix A, we describe alternatives we considered and detail how we decided which methods performed best. We restrict discussion to the best-performing approaches in the main paper.

Notation. Our metrics will assume indexed sets $\mathcal{D}$, $\mathcal{S}$ such that summary $S_i \in \mathcal{S}$ summarizes document $D_i \in \mathcal{D}$. The length in words of a sequence s is $|s|$ and the length in sentences is $\|s\|$. Each metric assigns a value $x \in [0, 1]$ to every (D_i, S_i) where 1 is the maximal score and example-level scores are averaged to yield a dataset-level score.

Compression. We quantify compression at the word (w) and sentence (s) levels:

$$\textbf{CMP}_w\left(D_i, S_i\right) = 1 - \frac{|S_i|}{|D_i|} \quad (1)$$

$$\textbf{CMP}_s\left(D_i, S_i\right) = 1 - \frac{\|S_i\|}{\|D_i\|} \quad (2)$$

Topic Similarity. We learn a topic model $\mathcal{M}$ on training corpus $\mathcal{T}$ with k topics using LDA (Blei et al., 2003) and quantify topic similarity by comparing the inferred topic distributions $\theta_{D_i|\mathcal{M}}, \theta_{S_i|\mathcal{M}}$ for a given summary and document:

$$\textbf{TS}\left(D_i, S_i\right) = 1 - \textsc{js}(\theta_{D_i|\mathcal{M}}, \theta_{S_i|\mathcal{M}}) \quad (3)$$

where JS is the Jensen-Shannon distance. We set $k = 20$ and $\mathcal{T} = \mathcal{D}$.

Abstractivity. Grusky et al. (2018) introduced *fragments* $\mathcal{F}(D_i, S_i)$, which are greedily-matched spans shared between D_i and S_i. We quantify abstractivity as a normalized function of the aggregate fragment length; our definition generalizes the definition of Grusky et al. (2018). We set $p = 1$.

$$\textbf{ABS}_p\left(D_i, S_i\right) = 1 - \frac{\sum_{f \in \mathcal{F}(D_i, S_i)} |f|^p}{|S_i|^p} \quad (4)$$

Redundancy. ROUGE (Lin, 2004) implicitly penalizes redundancy but underestimates its detrimental impacts (Chaganty et al., 2018). However, we find that ROUGE is effective for identifying redundancy given the definitional focus on overlapping spans. We quantify redundancy as the average ROUGE-L F-score for all pairs of distinct sentences in the summary.

$$\textbf{RED}\left(S_i\right) = \underset{(x,y) \in S_i \times S_i, x \neq y}{\text{mean}} \textsc{rouge}\left(x, y\right) \quad (5)$$

Semantic Coherence. We evaluate the semantic coherence of multi-sentence summaries by predicting the probability of each successive sentence conditioned on the previous one using a powerful language model, BERT (Devlin et al., 2019), pretrained with precisely this objective.

$$\textbf{SC}\left(S_i\right) = \frac{\sum_{j=2}^{||S||} \mathbb{1}_{\textsc{bert}}\left(S_i^j \mid S_i^{j-1}\right)}{||S_i|| - 1} \quad (6)$$

4 Data

We study the following 10 summarization datasets that have been frequently used in recent years.[6] Table 1 contains standard dataset statistics in the upper half and our aspect-level scores in the lower half; datasets are grouped by domain.

CNN-DM (Hermann et al., 2015; Nallapati et al., 2016) is a dataset composed of CNN and Daily Mail news articles with summaries that are a concatenated list of highlight bullet points.

NYT (Sandhaus, 2008) is a dataset of curated New York Times articles paired with abstracts written by library scientists.

NWS (Grusky et al., 2018) is the Newsroom dataset of news articles drawn from 38 top English publishers paired with multi-sentence summaries written by the original authors and editors.

GW (Graff and Cieri, 2003) is the Gigaword *headline generation* dataset that some refer to as a summarization dataset (Rush et al., 2015; Chopra et al., 2016). Examples in the dataset are drawn from seven news sources and are the article prefix paired with its headline.

XSum (Narayan et al., 2018) is an *extreme summarization* dataset where BBC articles are paired with single-sentence summaries written generally by the author of the article that tries to motivate the BBC audience to read the article by answering "What is the article about?".

PeerRead (Kang et al., 2018) is a dataset of paper drafts from top-tier computer science venues as well as arXiv.[7] Consistent with its use in the summarization community, we consider the full introduction to be the source document and the abstract to be the target summary.

[5] Different names and interpretations have been given for these properties in the literature. We revisit this in Appendix A in discussing alternate metrics.

[6] Several of these datasets are catalogued in the repository of Dernoncourt et al. (2018).

[7] Some papers also have peer reviews which we ignore.

PubMed (Cohan et al., 2018) is a dataset of papers drawn from the biomedical and life sciences. Unlike **PeerRead**, the full paper is taken as the document but the summary is still specified as the abstract.

TL;DR (Völske et al., 2017) is a dataset of user-written articles from the social media platform Reddit along with the author-provided courtesy summaries that tend to be multi-sentence. Völske et al. (2017) applied a series of preprocessing procedures to filter out bot-generated content.

AMI (Carletta et al., 2005) is a dataset of transcribed meetings, some which are naturally occurring and the rest of which are elicited, with hand-annotated summaries. The transcription process has multiple steps that are described extensively by Carletta et al. (2005). Various additional data provided within the **AMI** dataset is neglected in this work.

MovieScript (Gorinski and Lapata, 2015) is a dataset of movie scripts drawn from the Script-Base corpus that are aligned with user-written summaries sourced either from Wikipedia or IMDB. Various additional data provided within the **MovieScript** dataset is neglected in this work.

5 Results and Analysis

Compression scores quantitatively disambiguate summarization tasks.
Concretely, we observe **GW** has the lowest compression scores and while **GW** is sometimes described as a summarization dataset (Rush et al., 2015; Chopra et al., 2016), it is better seen as a *headline generation* dataset that is more in the style of *sentence compression* (as is suggested by $\|S_i\| = \|D_i\| = 1$). Conversely, **AMI** and **MovieScript** achieve the highest scores by a substantial margin and are *long-document summarization* datasets. Classifying new summarization datasets accurately may prove useful given that successful methods from one domain often do not extend to another and this shortcoming in generalization can be attributed to the differences in compression requirements (Cohan et al., 2018).

Given the goals stated in the XSum dataset paper, TL;DR may be a better choice than XSum.
In particular, Narayan et al. (2018) introduce **XSum** as a large dataset that legitimately requires abstraction. While **XSum** is more abstractive than other News datasets (barring **GW**) and is relatively large, **TL;DR** displays greater abstractivity, similar length summaries, and is 15 times larger. That said, Narayan et al. (2018) explore topic-oriented strategies in their work and such methods may be better suited to **XSum** given the **TS** scores.

CNN-DM and NYT are suboptimal for studying abstractive/extractive systems respectively.
Several recent works (See et al., 2017; Paulus et al., 2018; Li et al., 2018) have used **CNN-DM** to build and evaluate abstractive systems. Conversely, **NYT** has been used to build extractive systems (Hong and Nenkova, 2014; Li et al., 2016). Given our findings, we find both of these trends to be inconsistent with dataset properties and suboptimal given other preferable datasets for these purposes: **CNN-DM** is one of the least abstractive datasets and there are larger and more extractive alternatives to **NYT** such as **NWS**. Especially in the case of **CNN-DM**, we note that training learning-based systems (e.g. neural methods) using data with limited abstractivity implies the resulting summarizers will be limited in their ability to generate genuinely abstractive text. This is validated by empirical findings as both See et al. (2017) and Zhang et al. (2018) observe limited abstractivity in abstractive systems trained on **CNN-DM**. In light of this, we argue systems should be characterized as abstractive or not based on their empirical behavior rather than their theoretical capability.[8]

CNN-DM is not a representative benchmark for summarization as a whole.
Recent work (Kryscinski et al., 2019; Raffel et al., 2019) has explicitly portrayed **CNN-DM** as the benchmark dataset for summarization; the field has implicitly done this for several years (Kryscinski et al., 2019). While there is clear value in evaluating pretrained representations on summarization datasets, we caution against using **CNN-DM** as a stand-in for the entire summarization subfield. Instead, we suggest using a diverse group of datasets and not reducing a highly heterogeneous subfield to a single dataset. While this adds additional overhead, this cost is necessary to draw meaningful conclusions about the impact of advances on summarization broadly given the pronounced diversity in summarization datasets (Table 1).

Post-processing methods for mitigating redundancy may be needed for practical systems.
While evaluation on standard datasets using ROUGE may not penalize for this, redundancy is clearly un-

[8] Zhang et al. (2018) provide complementary arguments for this position.

	News					Scientific		Social Media	Meeting	Script
	CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
# ex.	287K	655K	995K	3804K	203K	9963	21K	3084K	97	1061
avg. $\lvert D_i \rvert$	717	822	677	34	438	1203	2394	238	6020	28K
avg. $\lvert S_i \rvert$	50	46	40	9.6	24	160	270	27	314	122
avg. $\lVert D_i \rVert$	31	34	26	1	19	54	95	11	568	3156
avg. $\lVert S_i \rVert$	3.52	1.00	1.75	1.00	1.00	6.10	10.0	1.71	17.1	5.14
CMP$_w$	0.909	0.869	0.910	0.714	0.904	0.763	0.870	0.876	0.941	0.994
CMP$_s$	0.838	0.915	0.890	0.001	0.902	0.765	0.874	0.811	0.964	0.998
TS	0.634	0.586	0.539	0.478	0.578	0.702	0.774	0.438	0.573	0.547
ABS$_1$	0.135	0.249	0.191	0.334	0.346	0.201	0.122	0.384	0.184	0.147
RED	0.157	-	0.037	-	-	0.168	0.17	0.056	0.215	0.152
SC	0.964	-	0.981	-	-	0.994	0.990	0.961	0.968	0.983

Table 1: **Upper half**: Standard dataset statistics. **Lower half**: Aspect-level scores for each dataset (0 is minimal value, 1 is maximal value). Corresponding standard deviations appear in Table 9. Redundancy and semantic coherence are not reported for datasets with $> 95\%$ single-sentence summaries.

desirable (Carbonell and Goldstein, 1998; Peyrard, 2019a) and existing datasets (and thereby systems learned using that data) display significant amounts of redundancy in their gold-standard summaries (exceptions are datasets with short summaries where cross-sentence redundancy is constrained to be low). Specifically, Nenkova (2006) argues that redundancy is a clear inhibitor for practical application of summarization systems. Consequently, *post hoc* methods that reduce redundancy after initial evaluation may be useful in generating summaries that are suitable for human users.

Semantic coherence captures observable variation in summary coherence.

We observe that the Scientific summaries (which are abstracts of published papers) are clearly more coherent than the author-generated summaries in **TL;DR**, the fragmented summaries in **AMI**, and the concatenated bullet-point summaries in **CNN-DM**. We find that this distinction is captured by the **SC** measure using BERT. Quantifying semantic coherence is especially important given that the coherence of reference summaries will inform the coherence of system summaries, especially for learning-based approaches. Akin to what we discuss for abstractivity, See et al. (2017) and Paulus et al. (2018) both demonstrate that neural summarizers generate incoherent summaries despite achieving high ROUGE scores.

5.1 Pairwise Correlations

While the properties we evaluate for do not exhaust all aspects of summarization that may be of interest, it is unclear to what extent different measures overlap in judgments. To quantify this, in Table 2 we report pairwise correlations for every pair of metrics. In each case, the value reported is the Spearman rank correlation coefficient ρ computed between the length 10 vectors containing the scores for each dataset.[9] $\rho = 1$ indicates perfect positive correlation (which is why we see this for all diagonal entries) and $\rho < 0$ indicates the metrics are anti-correlated.

	CMP$_w$	**CMP**$_s$	**TS**	**ABS**$_1$	**RED**	**SC**
CMP$_w$	1	0.733	-0.188	-0.406	-0.179	-0.321
CMP$_s$	0.733	1	0.042	-0.297	0.036	0.0
TS	-0.188	0.042	1	-0.564	0.75	0.643
ABS$_1$	-0.406	-0.297	-0.564	1	-0.429	-0.214
RED	-0.179	0.036	0.75	-0.429	1	0.321
SC	-0.321	0.0	0.643	-0.214	0.321	1

Table 2: Pairwise correlations measured using Spearman ρ coefficient between metrics studied in this work.

Unsurprisingly, the compression metrics are strongly correlated with each other. We further observe that redundancy and topic similarity are correlated whereas abstractivity is anti-correlated with both. In particular, when summaries are considerably redundant, we qualitatively observe that the repeated content in the summary was both important and repeated in the context of the reference document. As a result, this may explain why redundancy and abstractivity are anti-correlated as this would suggest that highly redundant summaries are highly extractive. Additionally, since we measure topic similarity using LDA and unigram count statistics, it is not surprising that extractions may correlate with high topic similarity. In part, this may suggest a deficiency of our measure of topic similarity to accurately consider references to the

[9] We omit scores for datasets that do not have scores for a given metric.

same topic using substantially different words.

We also observe that semantic coherence patterns similarly to redundancy. In particular, while we find the semantic coherence scores are appropriate for most examples we manually inspected, this suggests that BERT relies upon word-level overlaps in making next-sentence judgments (similar to behaviors seen in other sentence-pair tasks such as natural language inference, c.f Gururangan et al., 2018)

6 Detecting Low Quality Examples

To complement our quantitative dataset-level analysis, we conduct a qualitative study of individual examples by examining outliers. For each (dataset, metric) pair, we sample 10 examples from both the top and bottom 10% of examples for that metric and in that dataset.

Since manually considering all of the 1080 examples was not feasible, we began by examining the sampled examples for topic similarity, redundancy, and semantic coherence. Our hypothesis was that example quality would positively correlate with coherence and topic similarity and negatively correlate with redundancy. We found this hypothesis to be validated by our observations as we found that examples with low coherence, low topic similarity, or high redundancy scores were generally low quality examples. Every example which we judged to be low quality demonstrated at least one of the following defects:

- The summary contains critical disfluencies that severely hinder accurate processing.[10]
- The summary excludes unambiguously critical information from the reference document.
- Crucial information in the summary does not appear in the reference document and is not general knowledge.
- Substantial fractions of the summary involve entities, relations, or events that are ambiguous and that we could not resolve from the summary alone. In particular, accurate interpretation of the summary would require also reading the reference document to resolve various coreferring expressions; the summary is not self-contained.[11]
- The summary is entirely inappropriate as a summary of the reference document. For example, the summary only discusses an event with no obvious relationship to the contents of the reference document.
- The summary includes an entire sentence or long phrase describing something that appears in the main document but that is clearly an auxiliary detail. We flagged examples as low quality due to this condition quite conservatively, only using it when we could come to no basis for why the sentence/phrase should appear in the summary.

On the other hand, we did not find any systematic defects in examples with high coherence, high topic similarity, or low redundancy scores. Instead, almost all of these examples were satisfactory.

For the remaining two properties (compression measured by $\textbf{CMP}_w$, abstractivity measured by $\textbf{ABS}_1$), we analyzed all of the associated 400 examples. What we observed is that many of these examples tended to be generically low quality and we quantify this in Table 3. Since this analysis may be difficult to replicate and involves subjective decisions about example quality, we comprehensively enumerate all example IDs we use in Table 8.

Table 4 shows a representative subset of the low quality examples we found in our analysis. We provide further examples in Appendix C and Figures 1–9.

Compression. Minimally compressed summaries in **NYT**, **NWS**, **TL;DR**, and **PubMed** often are supplementary information to the document rather than a summary of it; in some cases, we believe this is due to errors in alignment in dataset construction/release. On the other hand, heavily compressed summaries in **NWS** and **XSum** often are just category labels (e.g. *Sports*), in **TL;DR** are usually attention-grabbers, and in **NYT** are near-exact duplicates of reference documents, which themselves are letters to the editor.

[10] We invoked this condition fairly judiciously as we observed that the domain of summaries also could influence the fluency of summaries in terms of grammaticality. In particular, we unsurprisingly found that academic papers in the Science domain generally have highly grammatical summaries whereas the bullet-point summaries in **CNN-DM** and the author-written summaries in **TL;DR** often were ungrammatical but still sufficiently clear to be interpreted correctly.

[11] Many summaries drawn from the News domain have references that could be resolved by world knowledge or that could be reasonably understood using common sense knowledge. In these cases, while the summary is not fully self-contained, we did not judge them to be low quality. However, we expect that systems trained using these datasets would require knowledge beyond what is afforded by the reference document to accurately generate summaries of this type.

	News					Scientific		Social Media	Meeting	Script
	CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
CMP$_w$ ↑	50	50	70	60	30	10	10	80	0	10
ABS$_1$ ↑	40	30	70	50	50	70	50	80	0	10
CMP$_w$ ↓	20	50	40	10	40	70	20	30	0	10
ABS$_1$ ↓	30	10	30	0	50	10	0	50	0	10

Table 3: **Upper half**: Percent of examples sampled from the top (↑) 10% for the given metric that were low quality. **Lower half**: Percent of examples sampled from the bottom (↓) 10% for the given metric that were low quality.

Dataset	Metric	Document	Summary
TL;DR	**CMP**$_w$ ↑	Brodie (the dog) was neglected . . . health issues concerning his skin. . . .	Onions
PeerRead	**ABS**$_1$ ↑	a lógica é o estudo dos princípios e critéiros de inferência . . .	logic is the science of correct inferences . . .
NWS	**CMP**$_w$ ↓	© Telegraph Media Group Limited 2016	David Moyes has returned to former club Manchester United . . .
TL;DR	**ABS**$_1$ ↓	Let us, in the beginning, give a word of cordial praise to the . . .	Let us, in the beginning, give a word of cordial praise to the . . .

Table 4: Representative low quality examples in the given dataset from the top (↑) or bottom (↓) 10% of examples for the given metric. Due to space constraints, some examples are abridged and shorter examples were preferred in selecting representatives. Additional examples are provided in Appendix C and Figures 1–9.

Abstractivity. Manual inspection reveals highly abstractive summaries in **NYT** and **NWS** generally are exceedingly vague or are entirely unrelated to the original document. Highly abstractive summaries in **PeerRead** are often translated to English from the reference document's language and discuss results that do not appear in the introduction but likely appear later in the paper. Conversely, extremely extractive summaries in **NWS** and **NYT** often are just the lede and cannot be understood without the reference document. However, in most other instances, the lede is an effective summary for examples drawn from the News domain.

Within the context of our sample of examples, we find that eight of the ten summarization datasets (all but **AMI**, **MovieScript**) contain at least 8% low quality examples, the majority contain at least 14% low quality examples, and that these low quality examples can be detected using our *compression* and *abstractivity* metrics. For the worst-offending **TL;DR** dataset, we conservatively estimate at least 20% of examples are of substantially subpar quality. In general, we find that the low quality **TL;DR** "summaries" we detect often serve a different rhetorical purpose than summarization (e.g. attention grabbing, responding to a previous post that is not available in the dataset, sarcasm/humor).

7 Related Work

Dataset Analysis. As an alternative to automated evaluation, Chen et al. (2016) and Yatskar (2019) conduct human evaluations of standard datasets in reading comprehension and question answering. In some cases, dataset creators perform manual analyses of the data they introduce (e.g. Sandhaus (2008) and Grusky et al. (2018) for the **NYT** and **Newsroom** corpora, respectively). Automated and human evaluation provide complementary benefits with respect to their scalability and reliability. Even in the context of human evaluations, we advocate that automatic metrics can be useful in guiding the exploration of data and informing subsampling procedures that provide fine-grained insights.

Quality Estimation. Our work bears resemblance both in name and structure to work on quality estimation. Quality estimation, often centered on natural language generation, is the task of measuring system-generated output quality (Paetzold and Specia, 2016; Yuan and Sharoff, 2020). It is closely related to work on unsupervised or reference-free evaluation (Napoles et al., 2016; Ethayarajh and Sadigh, 2020). Within the context of summarization, the special case of quality estimation regarding factual consistency/faithfulness has been of recent interest (Wang et al., 2020; Maynez et al., 2020; Durmus et al., 2020) since neural abstractive summarizers have been shown to hallucinate/misrepresent facts (See et al., 2017). In comparison to these settings, our metrics make no use of labelled data (even in training) and are entirely intrinsic/unsupervised.

Summarization Practices. Several analyses and critiques exist for different aspects of the summarization pipeline. From a modelling perspective, Zhang et al. (2018) assess whether abstractive systems are truly abstractive, Kedzie et al. (2018) evaluate content selection policies in a variety of

methods, and Mao et al. (2020) assess the facet-level performance of extractive summarizers. From an evaluation perspective, several works have discussed the shortcomings of ROUGE/automated evaluation (Liu and Liu, 2008; Chaganty et al., 2018; Hashimoto et al., 2019; Peyrard, 2019b) as well proposed alternative metrics for summarization or natural language generation more broadly (Clark et al., 2019; Zhang et al., 2020; Sellam et al., 2020).

Two recent works are highly related to our own. Kryscinski et al. (2019) provide a critical reevaluation of summarization research. Most relevant to our work, they show that web-scraped datasets, specifically **CNN-DM** and **NWS**, contain a nontrivial fraction of examples (approx. 3.5%) with HTML artifacts (which can be easily detected/removed). Jung et al. (2019) provide an aspect-level evaluation of both summarization datasets and systems. In their work, the dataset analyses center on biases in the data (e.g. positional biases, which are often seen in news summarization), which is reminiscent of the annotation artifacts seen in other NLP tasks (Gururangan et al., 2018; Niven and Kao, 2019).

8 Discussion

Open Problems and Future Directions. Our results demonstrate that a sizeable fraction of examples in most summarization datasets are low quality. However, it remains open whether modellers should simply prune these examples, manually/automatically attempt to correct them, or model them without change. We do note that research in the machine learning and learning theory communities shows that models both theoretically and empirically do substantially worse when trained using low quality examples, even when the examples are not strictly adversarially chosen (Klivans et al., 2009; Biggio et al., 2012; Koh et al., 2018). These concerns are further compounded by the evidence of Belinkov and Bisk (2018) that neural models for natural language generation are not robust to naturally noisy data.

Our metrics may be repurposed to rank examples in designing curricula for curriculum learning approaches (Bengio et al., 2009). Alternatively, they can serve as additional metrics for the (possibly unsupervised) evaluation of summarization systems, potentially mitigating deficiencies in standard metrics, such as ROUGE, by directly penalizing redundancy and semantic incoherence.

Limitations. In this work, we restrict ourselves to single-document single-reference English language summarization datasets. While the datasets we study constitute a considerable fraction of dataset usage in the summarization community, several multi-document summarization datasets have been introduced (e.g. Fabbri et al., 2019; Antognini and Faltings, 2020) and multi-reference summarization datasets have often been argued to be desirable due to under-constrained nature of the summarization task (Kryscinski et al., 2019) and the ideal evaluation paradigm for ROUGE (Lin, 2004). Beyond English, both large summarization datasets (Nguyen and Daumé III, 2019; Varab and Schluter, 2020) and more general language resources/technologies (Joshi et al., 2020) are less available, which may heighten the need for data quality assurance.

More broadly, the measures that we introduce are automated, and therefore non-human, judgments of the quality of summarization data. Therefore, we only envision these measures to be useful as inexpensive first-order approximations of aspect-level summary quality rather than bona fide replacements for human evaluation. Additionally, since we principally envision applying these metrics to datasets, we make no efforts to make these metrics robust to adversarially-crafted data and they are likely quite susceptible to adversarial attack.

9 Conclusion

In this work, we demonstrate that various aspects of summarization datasets can be intrinsically evaluated for. We specifically show this for 5 properties across 10 popular datasets, uncovering that dataset use is sometimes incongruous with the attributes of the underlying data. We also find that some aspect-level estimators may be surprisingly effective at detecting low quality dataset examples. Our findings suggest that more intentional and deliberate decisions should be made in selecting summarization datasets for downstream modelling research and that further scrutiny should be placed upon summarization datasets released in the future.

10 Reproducibility

All code is made publicly available.[12] Exhaustive reproducibility details, including how to access all datasets, are provided in Appendix B. We fully

[12] https://github.com/rishibommasani/SummarizationEvaluationEMNLP2020

adhere to the EMNLP 2020 Reproducibility guidelines, addressing all relevant checklist items.

Acknowledgments

We thank Anna Huang for her help with analyzing the data. We thank Ge Gao, Esin Durmus, and members of the Cornell and Stanford NLP groups for their valuable advice. We especially thank the reviewers and area chairs for their articulate and constructive feedback.

References

Diego Antognini and Boi Faltings. 2020. GameWikiSum: a novel large multi-document summarization dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6645–6650, Marseille, France. European Language Resources Association.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pages 41–48, New York, NY, USA. Association for Computing Machinery.

Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1467–1474, Madison, WI, USA. Omnipress.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Pickman Mann, Nick Ryder, Melanie Subbiah, Jean Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric J Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Chris Callison-Burch, Lyle Ungar, and Ellie Pavlick. 2015. Crowdsourcing for NLP. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–3, Denver, Colorado. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

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, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Franck Dernoncourt, Mohammad Ghassemi, and Walter Chang. 2018. A repository of corpora for summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John C. Duchi. 2019. Lecture notes for statistics 311/electrical engineering 377.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

H. P. Edmundson. 1969. New methods in automatic extracting. *J. ACM*, 16(2):264–285.

Kawin Ethayarajh and Dorsa Sadigh. 2020. Bleu neighbors: A reference-less approach to automatic evaluation. *ArXiv*, abs/2004.12726.

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.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. *ArXiv*, abs/1803.09010.

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.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Denver, Colorado. Association for Computational Linguistics.

David Graff and Christopher Cieri. 2003. English Gigaword. Philadelphia, Pennsylvania. Linguistic Data Consortium.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. 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 C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Matthew D. Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pages 856–864, USA. Curran Associates Inc.

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*, pages 712–721, Gothenburg, Sweden. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. 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 3322–3333, Hong Kong, China. Association for Computational Linguistics.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. 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.

Adam R. Klivans, Philip M. Long, and Rocco A. Servedio. 2009. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(94):2715–2740.

Pang Wei Koh, Jacob Steinhardt, and Percy Liang. 2018. Stronger data poisoning attacks break data sanitization defenses. *ArXiv*, abs/1811.00741.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018. Improving neural abstractive document summarization with structural regularization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4078–4087, Brussels, Belgium. Association for Computational Linguistics.

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.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Inderjeet Mani. 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA.

Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han. 2020. Facet-aware evaluation for extractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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*, Seattle, Washington. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. ACM.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, 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.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova. 2006. *Understanding the Process of Multi-document Summarization: Content Selection, Rewriting and Evaluation*. Ph.D. thesis, New York, NY, USA. AAI3203761.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer.

Khanh Nguyen and Hal Daumé III. 2019. Global voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. SimpleNets: Quality estimation with resource-light neural networks. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 812–818, Berlin, Germany. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Maxime Peyrard. 2019a. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard. 2019b. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*,

pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W. Black. 2020. Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

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*, abs/1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. Philadelphia, Pennsylvania. Linguistic Data Consortium.

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, Valencia, Spain. 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.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

Karen Spärck Jones. 1999. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc.

Daniel Varab and Natalie Schluter. 2020. DaNewsroom: A large-scale Danish summarisation dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6731–6739, Marseille, France. European Language Resources Association.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Rui Yan, Mingkun Gao, Ellie Pavlick, and Chris Callison-Burch. 2014. Are two heads better than

one? crowdsourced translation via a two-step collaboration of non-professional translators and editors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1134–1144, Baltimore, Maryland. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc.

Mark Yatskar. 2019. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2318–2323, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Yuan and Serge Sharoff. 2020. Sentence level human translation quality estimation with attention-based neural networks. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1858–1865, Marseille, France. European Language Resources Association.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.

Fangfang Zhang, Jin-ge Yao, and Rui Yan. 2018. On the abstractiveness of neural document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 785–790, Brussels, Belgium. Association for Computational Linguistics.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022.

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*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

A Alternative Metrics

A.1 Compression

For compression, we found sentence-level compression to be a naturally motivated metric given that many extractive systems are constrained to extract sentence-length sequence. We also considered byte-level compression as an alternative to word-level compression (as computational length constraints have sometimes been used in evaluation instead of word length constraints). We found the results to be highly correlated with word-level compression and to not be further revealing (and bytes may be inherently less interpretable for NLP when compared with words). We also considered only considering content words, motivated by literature in topic modelling (Schofield et al., 2017) that has considered removing stopwords and other such lexical categories. These results were also highly correlated with the original word-level compression results and we did not find any discerning trends in looking at individual examples.

A.2 Topic Similarity

In the main paper, we compute topic similarity using the Jensen-Shannon distance. We initially considered the Kullback-Leibler (KL) divergence. While the JS distance and/or divergence has been more frequently used in the context of similarity in topic modelling, the KL divergence is also frequently considered. Intuitively and under some interpretations, the asymmetry of the KL divergence may be desirable as the extent to which a summary is topically similar to a document may not be the same as the extent to which a document is topically similar to a summary. In spite of this, in viewing the results using KL, we found that the measure lacked discriminative power in disambiguating examples we believed were more topically similar than others. We qualitatively found the judgments via the JS distance to be accurate. That said, the judgments between the measures tended to be highly correlated as the Spearman rank correlation coefficient was $\rho \geq 0.7$ for all topic modelling settings and in most cases exceeded 0.8.

We also considered a topic model learned using both the documents and summaries $\mathcal{D} \cup \mathcal{S}$ and just the documents $\mathcal{D}$. Both are natural choices, with using the documents being more general in some sense as the topic similarity of a summary should be able to be assigned without requiring the summary collection. We further considered several

choices for the number of topics as well. In Table 5, we report the full results for all pairs of (training corpus $\mathcal{T}$, # topics k) for all

$$(\mathcal{T}, k) \in \{\mathcal{D} \cup \mathcal{S}, \mathcal{D}\} \times \{10, 20, 50, 100\}.$$

In all cases, the number of training examples is truncated to 20000 (hence 10000 summaries and 10000 documents when using the training corpus of $\mathcal{D} \cup \mathcal{S}$). We fix the number of training documents across datasets to attempt to control for the confound of larger datasets inducing higher quality topic models. We did not observe significant changes in the result by relaxing this (i.e. using the full datasets instead of just 20000 examples).

We find that there is significant variation in cross-dataset rankings with respect to these two parameters. We chose to report the results corresponding to $k = 20, \mathcal{T} = \mathcal{D}$. We chose the value for k based on qualitative judgments about topic quality for **CNN-DM**, **PeerRead**, and **AMI**, as we considered these to be a diverse subset of all 10 datasets. The topics we observed were highly disjoint and reasonably aligned with our intuitions about what sensible topics should be. We chose the value for $\mathcal{T}$ based on the generality referenced previously. While the results are substantially different for $\mathcal{D}$ versus $\mathcal{D} \cup \mathcal{S}$, we did not find any consistent and interpretable discriminative properties between the two.

A.3 Abstractivity

Our general framework for quantifying abstractivity is derived from Grusky et al. (2018). We considered $p \in \{1, 2, 3, 4\}$ initially and found $p = 1$ to be the most informative regarding abstractivity. In particular, we find that for increasing p, useful conclusions about abstractivity are inherently masked by the dominance of the $|S_i|^p$ denominator in the definition. We report the scores for $\mathbf{ABS}_2$ in Table 6.

We also considered the natural extensions to $\mathbf{ABS}_3$ and $\mathbf{ABS}_4$ but we found that the normalization dominates any deviation in the scores and all datasets essentially receive a score of 1. We also considered other forms of normalization (i.e. normalizing $\mathbf{ABS}_2$ in the style of the L_2 norm/the style of generalized p-norms) in initial experiments but found no substantial differences.

A.4 Redundancy

In the main paper, we compute redundancy scores for each distinct sentence pair using ROUGE-L F-measure and then average these individual values to get a score for the entire summary. Alternatively, we considered other ROUGE scores (specifically ROUGE-1 and ROUGE-2) as well as max pooling the sentence pair scores. We report these results below in Table 7.

We do not observe significant changes with the specific ROUGE metric considered (i.e. a Spearman ρ of 1.0 which indicates a perfect correlation in the case of max pooling across the ROUGE variants). We do see substantial differences between averaging and max pooling; we find that max pooling turns out to precisely correlate ($\rho = 1.0$) with the average summary length. This is somewhat expected, given that the max-pooled redundancy estimates doesn't inherently control for summary length. We therefore chose to report redundancy scores using averaging as we also qualitatively found them to be more useful and characteristic, especially for datasets such as **AMI** and the Scientific datasets as max pooling was overly aggressive. While the nuances of the specific ROUGE variant did not significantly impact trends in redundancy scores, we chose to report the ROUGE-L scores in the main paper as we (highly subjectively) found the values to be most interpretable/consistent with values we would have assigned.

A.5 Semantic Coherence

We evaluate for semantic coherence between successive pairs of sentences, exploiting the auxiliary training objective of BERT beyond its masked language modeling objective. In particular, we were especially interested in this given that many systems are designed with explicit handling of sentence boundaries (e.g. more extractive systems first rank extractive sentences and then order a thresholded subset) and datasets such as **CNN-DM**, which are artificially concatenated, may not be inherently coherent across sentence-boundaries.

Our observations regarding the measure of coherence provided by BERT's next-sentence predictions seem to contradict existing findings. In particular, Liu et al. (2019) introduce RoBERTa as a direct followup study to BERT and find that the next-sentence prediction objective is not an effective pretraining objective for improving representations for natural language understanding; Yang et al. (2019) also provide similar evidence. However, our findings do not contest these conclusions but instead suggest that, nonetheless, BERT

k	$\mathcal{T}$	News					Scientific		Social Media	Meeting	Script
		CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
10	$\mathcal{D}$	0.715	0.666	0.616	0.546	0.629	0.769	0.812	0.536	0.702	0.553
10	$\mathcal{D} \cup \mathcal{S}$	0.805	0.81	0.809	0.864	0.8	0.854	0.835	0.847	0.332	0.613
20	$\mathcal{D}$	0.634	0.586	0.539	0.478	0.578	0.702	0.774	0.438	0.573	0.547
20	$\mathcal{D} \cup \mathcal{S}$	0.773	0.757	0.771	0.87	0.763	0.815	0.751	0.823	0.361	0.463
50	$\mathcal{D}$	0.572	0.507	0.472	0.414	0.497	0.64	0.721	0.368	0.561	0.445
50	$\mathcal{D} \cup \mathcal{S}$	0.708	0.694	0.705	0.769	0.693	0.752	0.698	0.71	0.347	0.411
100	$\mathcal{D}$	0.519	0.468	0.416	0.385	0.422	0.601	0.679	0.318	0.536	0.432
100	$\mathcal{D} \cup \mathcal{S}$	0.681	0.66	0.665	0.689	0.632	0.725	0.667	0.638	0.35	0.395

Table 5: Alternative methods for estimating redundancy. Results in main paper are equivalent to those in the row corresponding to 20 and $\mathcal{D}$.

	News					Scientific		Social Media	Meeting	Script
	CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
$\mathbf{ABS}_1$	0.135	0.249	0.191	0.334	0.346	0.201	0.122	0.384	0.184	0.147
$\mathbf{ABS}_2$	0.932	0.917	0.762	0.862	0.953	0.943	0.983	0.932	0.995	0.983

Table 6: Alternative methods for estimating abstractivity. Results in the main paper are for $\mathbf{ABS}_1$.

is a strong next-sentence predictor and that these predictions are still useful for measuring coherence across sentences. While we considered word or subword measures of coherence, we did not consider alternative pretrained models that are pretrained on other objectives related to inter-sentence coherence such as ALBERT (Lan et al., 2020). Given the findings of Lan et al. (2020, §4.6), it seems likely that the sentence order prediction task they use may be more effective for measuring semantic coherence. Concurrent work by Prabhumoye et al. (2020) also substantiates the usefulness of BERT-based next-sentence prediction for measuring coherence and ranking sentences orders.

That said, semantic coherence could also be evaluated using (neural) language models, especially in light of results suggest they may be consistent with human judgments regarding grammaticality and acceptability (Chowdhury and Zamparelli, 2018; Warstadt et al., 2019). We did consider this and found language modeling scores (e.g. surprisal) assigned via a pretrained high-quality causal language model (GPT-2) to be inconsistent with our human judgments. We believe language modeling scores in this sense are likely highly sensitive to the domain (and even within-domain effects, e.g. lexical variation for **XSum** which is fairly limited given all articles are sourced from the BBC whereas for **Newsroom** the variation is greater given the heterogeneous group of publishers with more diversified writing styles).

B Reproducibility Details

We provide precise and comprehensive details discussing all data, preprocessing and modelling decisions. All code will be made publicly available as noted in the main paper.

B.1 Dataset Sources

We use the versions of **GW** and **CNN-DM** dataset released by Gehrmann et al. (2018).[13] Sentence boundary tokens inserted by Gehrmann et al. (2018) to improve summarization quality were removed to ensure fair comparison in our work. An important distinction in the use of the **CNN-DM** dataset for modeling is whether the entity-anonymized or non-anonymized version was used. This copy is non-anonymized and it is important to consider the stability of our metrics under this anonymization.
We used the released version of the **NYT** dataset directly as it was released via LDC.[14]
We use the released version of the **TL;DR** dataset provided by the authors of Völske et al. (2017).[15]
We use a version of the **NWS** dataset that was released via private communication with the authors of Grusky et al. (2018). We have verified with the authors that the data can be requested with the platform they released in their original work.[16]
For all remaining datasets, we use the version re-

[13] https://github.com/harvardnlp/sent-summary
[14] https://catalog.ldc.upenn.edu/LDC2008T19
[15] https://tldr.webis.de/
[16] https://summari.es/

		News					Scientific		Social Media	Meeting	Script
		CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
max	ROUGE-1	0.266	-	0.067	-	-	0.36	0.457	0.082	0.635	0.292
max	ROUGE-2	0.049	-	0.014	-	-	0.123	0.225	0.014	0.453	0.062
max	ROUGE-L	0.238	-	0.055	-	-	0.287	0.385	0.074	0.616	0.227
mean	ROUGE-1	0.172	-	0.045	-	-	0.214	0.215	0.063	0.239	0.195
mean	ROUGE-2	0.014	-	0.004	-	-	0.027	0.033	0.006	0.041	0.015
mean	ROUGE-L	0.157	-	0.037	-	-	0.168	0.17	0.056	0.215	0.152

Table 7: Alternative methods for estimating redundancy. Results in main paper are equivalent to those in the row corresponding to mean and ROUGE-L.

leased by Jung et al. (2019).[17] All of our conventions in using these five datasets follow their work.

B.2 Data Preprocessing

All datasets were first filtered to remove examples where either the document or summary was empty. We found only examples in **CNN-DM** failed this criterion and this constituted less than 0.1% $\left(\frac{114}{287227}\right)$ of the dataset.

All results were reported then on the standard training set if we were aware of a standard split used consistently in the summarization system literature. Splits in the case of datasets sourced from the work of Jung et al. (2019) followed their work. In all cases, the training set was at least 80% of the full data collection, so we expect results to generalize to the portions of the collection that were not considered assuming splits were constructed by sampling uniformly at random (we did not verify this).

Sentence-level tokenization was performed using `NLTK` (Loper and Bird, 2002). Word-level tokenization was performed using `SpaCy` (Honnibal and Montani, 2017).

B.3 Topic Similarity

We lowercase all terms, remove stopwords using the list specified in `NLTK` (Loper and Bird, 2002), and lemmatize using `SpaCy` (Honnibal and Montani, 2017). We only retain words tagged with a POS category in $\{$*NOUN, ADJ, VERB, ADV*$\}$ by the `SpaCy` POS tagger. We use LDA (Blei et al., 2003) to learn all topic models and rely on the implementation in `Gensim` (Řehůřek and Sojka, 2010) based on specification of Hoffman et al. (2010). All hyperparameters are set as default and we discussed the number of topics k and training corpus $\mathcal{T}$ in §A.2 with the results in the main paper using $k = 20$ and $\mathcal{T} = \mathcal{D}$ where $\mathcal{T}$ is truncated to be at most 20000 documents. We compute the Jensen-Shannon distance using `SciPy` (Virtanen et al., 2020).

B.4 Abstractivity

Fragments (Grusky et al., 2018) were computed using the scripts released in that work for the purposes of estimating abstractivity. In the case of the **NWS** dataset, the authors already provide fragment-related scores which we use without recomputing these values.

B.5 Redundancy

We make use of the native Python reimplementation of ROUGE (Lin, 2004), `easy-rouge`.[18] All scores reported in the main paper use ROUGE-L and use the computed F-measure score.

B.6 Semantic Coherence

We compute semantic coherence by predicting the probability of a sentence conditional on the preceding sentence using BERT. BERT was pretrained with exactly this objective (beyond its masked language modeling objective) and we use the released model as-is with no further fine-tuning. We use the `bert-base-uncased` model along with the associated tokenizer that was implemented in `PyTorch` (Paszke et al., 2017) by HuggingFace in the `transformers` repository.[19]

B.7 Efficiency

All metrics reported in the main paper can be computed over all datasets in less than 10 ten hours on a single CPU. The only model with a nontrivial number of parameters used in this work is the `bert-base-uncased` models we use in measuring semantic coherence. We refer readers to Devlin et al. (2019) for more details and to the

[17] http://biassum.com/

[18] https://github.com/neural-dialogue-metrics/rouge

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

HuggingFace implementation we reference previously.

C Detecting Low-Quality Examples

In the main paper, we briefly discuss how we discovered that several of our metrics can serve the dual purpose of detecting generally low quality examples for example that achieve extreme scores. Figures 1 through 9 are several examples we found to be representative of the general structure of low quality examples for a given metric. In some cases, the trends are highly dataset-specific whereas in others they are more general. To facilitate reproducibility efforts, we provide all examples IDs we studied for each (dataset, metric) in Table 8.

Original Text (truncated): Let us, in the beginning, give a word of cordial praise to the American publishers of these splendid volumes. The undertaking, in the first place, was an intellectual compliment to the country. It was based on the faith that there is in this country enough of philosophy and scholarship to justify a new and complete edition of . . .

Summary: Let us, in the beginning, give a word of cordial praise to the American publishers of these splendid volumes. The undertaking, in the first place, was an intellectual compliment to the country.

Detector: Extremely Low Abstraction

Figure 1: **Dataset: NWS**. This summary simply is the lede and we do not find it to be a useful summary for readers not familiar with the full context of the article. We hypothesize that such a summary may have been useful for members of a newsroom communicating information about the article to the other (given their intimate familiarity with the article) but this likely is inappropriate as a summary in most settings.

Original Text (truncated): A FULL-SERVICE hotel and conference center is to go up in the Lafayette Yard area of Trenton, giving the city a hotel for the first time since the 1980's and bringing to an end its unenviable distinction as the only state capital without lodging for visitors . . .

Summary: Acquest

Detector: Extremely Low Abstraction

Figure 2: **Dataset: NYT**. This summary simply conveys no useful information to someone who has not also read the reference document and simply is a word copied from the source document. It appears to be a label rather than a summary.

Original Text (truncated): a lógica é o estudo dos princípios e critéiros de inferências e demonstrações válidas. um sistema lógico é composto por três partes: a sintaxe (ou notação), . . .

Summary (truncated): logic is the science of correct inferences and a logical system is a tool to prove assertions in a certain logic in a correct way . . .

Detector: Extremely High Abstraction

Figure 3: **Dataset: PeerRead**. This summary simply is not in the same language and hence achieves a very high abstractivity.

Original Text (truncated): from russia with love"screenplay byrichard maibaumadapted byjohanna harwoodbased on the novel byian fleming . . .

Summary: final

Detector: Extremely High Abstraction

Figure 4: **Dataset: MovieScript**. This summary simply bears no clear relationship with the reference document and therefore repeats no words and achieves maximal abstractivity.

Original Text: BASEBALL American League BALTIMORE ORIOLES – Agreed to terms with INF-OF Mark McLemore on a minor league contract. BOSTON RED SOX – Named Dale Sveum third base coach.

Summary: Sports transactions

Detector: Extremely High Abstraction

Figure 5: **Dataset: NYT**. This summary is unlikely to be informative to someone who has not read the reference document and is more of a categorization/label than a summary. This is similar to the previous **NYT** example given.

Original Text: © Telegraph Media Group Limited 2016

Summary: David Moyes has returned to former club Manchester United to strengthen his Sunderland squad after agreeing a fee for Paddy McNair and Donald Love.

Detector: Extremely Low Compression

Figure 6: **Dataset: NWS**. This summary has a negative compression score and, in this case, this seems to indicate the summaries and documents were extracted inaccurately using the scraper of Grusky et al. (2018).

D Mutual Information Bounds

The *entropy* of a random variable X is defined as:

$$H(X) \triangleq -\sum_{x} p(x) \log_2 p(x)$$

CNN-DM	**CMP**$_w$ $\uparrow$	8519	2640	5785	942	17538	7161	13516	19330	16770	8112
CNN-DM	**CMP**$_w$ $\downarrow$	4390	18955	14330	7336	17247	2380	13721	1560	16593	13157
CNN-DM	**ABS**$_1$ $\uparrow$	10483	4788	10191	1785	15750	17503	18399	13140	6154	7871
CNN-DM	**ABS**$_1$ $\downarrow$	15918	10958	16845	15301	18909	17897	13862	9637	8617	10269
NYT	**CMP**$_w$ $\uparrow$	11096	15782	14059	4182	266	5973	9748	17554	4002	3736
NYT	**CMP**$_w$ $\downarrow$	18308	10972	15081	16664	12310	7184	1692	4635	2783	18409
NYT	**ABS**$_1$ $\uparrow$	17019	11500	15663	15056	9464	5355	15736	13315	13404	15687
NYT	**ABS**$_1$ $\downarrow$	12317	6821	13615	6220	17242	18480	6280	3808	16364	5825
NWS	**CMP**$_w$ $\uparrow$	6627	507	4999	19020	10546	5215	11450	8467	19640	5027
NWS	**CMP**$_w$ $\downarrow$	12213	18094	11644	11969	3595	67	13752	12180	7927	4137
NWS	**ABS**$_1$ $\uparrow$	16092	19307	7422	6358	2191	17874	13484	16894	18728	4671
NWS	**ABS**$_1$ $\downarrow$	10698	1172	3014	9373	688	5724	7391	10575	1841	16314
TL;DR	**CMP**$_w$ $\uparrow$	15659	7458	9830	18016	435	15820	926	8790	12533	9555
TL;DR	**CMP**$_w$ $\downarrow$	7313	9667	12707	5431	19761	1577	10484	18118	15612	9623
TL;DR	**ABS**$_1$ $\uparrow$	15252	14719	3623	18758	6311	9860	12394	11822	12873	2787
TL;DR	**ABS**$_1$ $\downarrow$	4048	5538	18552	9621	4059	2044	1756	1927	906	12768
GW	**CMP**$_w$ $\uparrow$	6479	1795	9370	2274	11622	8430	6808	18236	7909	4108
GW	**CMP**$_w$ $\downarrow$	9276	3375	10192	2434	1471	12854	10455	13995	10361	5945
GW	**ABS**$_1$ $\uparrow$	3358	13215	2592	19244	16380	15535	10255	8373	15101	3056
GW	**ABS**$_1$ $\downarrow$	11466	5816	16528	11168	7642	10496	14	8223	13731	4971
AMI	**CMP**$_w$ $\uparrow$	96	92	18	62	0	28	74	51	45	
AMI	**CMP**$_w$ $\downarrow$	4	11	25	84	33	42	94	64	49	
AMI	**ABS**$_1$ $\uparrow$	43	49	25	10	28	29	41	74	42	
AMI	**ABS**$_1$ $\downarrow$	63	91	37	67	79	70	54	48	35	
MovieScript	**CMP**$_w$ $\uparrow$	979	393	185	140	977	186	335	567	688	399
MovieScript	**CMP**$_w$ $\downarrow$	159	343	133	693	896	14	1050	23	838	744
MovieScript	**ABS**$_1$ $\uparrow$	659	783	994	941	980	796	1060	207	86	338
MovieScript	**ABS**$_1$ $\downarrow$	445	488	253	733	233	158	978	391	553	341
PeerRead	**CMP**$_w$ $\uparrow$	358	744	54	9520	703	1629	4066	7122	2573	5711
PeerRead	**CMP**$_w$ $\downarrow$	3433	1877	757	1621	8257	7654	3635	3302	3807	5495
PeerRead	**ABS**$_1$ $\uparrow$	9128	4204	7638	3729	3354	3747	2614	6485	2533	6082
PeerRead	**ABS**$_1$ $\downarrow$	2910	1120	2157	212	9765	583	5653	48	729	6418
PubMed	**CMP**$_w$ $\uparrow$	9769	11434	19055	10724	5961	13804	4846	16193	11958	9084
PubMed	**CMP**$_w$ $\downarrow$	6335	7884	2919	17888	14458	13529	13062	18799	3435	5780
PubMed	**ABS**$_1$ $\uparrow$	5303	17763	4886	18555	17871	13251	5975	10611	14676	14655
PubMed	**ABS**$_1$ $\downarrow$	11705	2639	11863	5064	7551	530	1981	7509	8827	16006
XSum	**CMP**$_w$ $\uparrow$	18913	10476	11067	8546	2277	6992	3676	10926	4369	19607
XSum	**CMP**$_w$ $\downarrow$	164	16910	15343	12875	10730	15297	9999	14526	6751	7753
XSum	**ABS**$_1$ $\uparrow$	2942	14493	7669	12180	9360	19036	15122	12422	8353	660
XSum	**ABS**$_1$ $\downarrow$	3454	17269	11358	13847	18482	10213	10394	5319	15605	2627

Table 8: Exhaustive list of example IDs we studied in the evaluation described in Section 6 of the main paper. $\uparrow$ indicates the examples are sampled from the top 10% for a given metric, $\downarrow$ indicates the examples are sampled from the bottom 10% for a given metric. Since **AMI** has 97 summaries (which is less than 100), it is impossible to select 10 unique examples from either the top or bottom 10% for a given metric. Therefore, we simply consider the 9 examples within the top or bottom 10%.

Original Text: An article yesterday about plans by members of the House Intelligence Committee to visit Libya misidentified the member of Congress who headed a delegation to that country last month. He was Curt Weldon, Republican of Pennsylvania, not Tom Lantos, Democrat of California.

Summary: Six members of House Intelligence Committee are scheduled to meet in Libya with Col Muammar el-Qaddafi and other top Libyan officials, in second meeting between American Congressional delegation and Qaddafi since Libya agreed to dismantle its chemical and biological weapons program; members of House panel hope to use meeting to gauge accuracy of earlier American intelligence about Libya (M)

Detector: Extremely Low Compression

Figure 7: **Dataset: NYT**. Similar to the previous example, this summary has a negative compression score and, in this case, this seems to indicate the summaries and documents were created/aligned incorrectly in Sandhaus (2008).

Original Text (truncated): Brodie (the dog) was neglected, and ended up with serious anger and health issues concerning his skin and allergies. My boyfriend adopted him . . .

Summary: Onions.

Detector: Extremely High Compression

Figure 8: **Dataset: TL;DR**. We observe this trend quite frequently in **TL;DR**. Specifically, since authors on the social discussion platform Reddit choose to provide these summaries at their discretion, we often find the "summaries" are attention-grabbing and serve a starkly different rhetorical purpose from how summaries are generally conceived.

Original Text (truncated): these are external links and will open in a new window1908 - king carlos and eldest son assassinated in lisbon. second son manuel becomes king. 1910 - king manuel ii abdicates amid revolution . . .

Summary: a chronology of key events :

Detector: Extremely High Compression

Figure 9: **Dataset: XSum**. We observe this trend quite frequently in **XSum**. For articles that are essentially timelines or other types of chronologies discussing historic events diachronically (which forms a small but distinctive section of the writing style of BBC from our analysis), the summary extracted to accompany it is generally this string or a slightly altered version. We argue this summary is fairly unhelpful (and is likely fairly uninteresting to test models on; simple rule-based filtering made be preferable to avoid overestimating performance on this dataset because of these examples).

The *conditional entropy* of X given Y is defined as:

$$H(X \mid Y) \triangleq \sum_y p(y) \left[- \sum_x p(x \mid y) \log_2 p(x \mid y) \right]$$

The *mutual information* between random variables X and Y is defined as:

$$I(X;Y) \triangleq H(X) - H(X \mid Y)$$

The entropy measures the uncertainty in the probability mass/density function of a random variable. As such, the mutual information measures how much the entropy of X is reduced by (on average) due to the observation of Y.

In the main paper, we state the following inequality:

$$\underbrace{I(S;M)}_{\text{learned model}} \leq \underbrace{I(S;T)}_{\text{training data}} + \underbrace{I(S;P)}_{\text{pretraining}} + \underbrace{I(S;A)}_{\text{inductive bias}},$$

where I denotes the mutual information, S denotes understanding of the underlying summarization task and M denotes a model learned using summarization training data T, additional pretraining data P, and the model's architecture A.

Intuitively, the claim is that the uncertainty about the summarization task that is reduced by the model (which is uniquely determined by its training data, pretraining data, and architecture) is at most what can be cumulatively reduced by the training data, pretraining data, and inductive biases encoded in the model's architecture.

Our hypothesis is that $I(S;A)$ is small for learning-based models with minimal inductive biases, such as neural networks. Further, we hypothesize that while $I(S;P)$ is likely nontrivial for popular pretraining regimes, the dominant term on the right-hand side is likely $I(S;T)$. We do note that this second hypothesis may be false given the partial evidence of GPT-3 (Brown et al., 2020) and the successes it enjoys in few-shot learning due to pretraining at unprecedented scale. However, no evaluation is conducted on summarization data in that work.

Proof.

$$\begin{aligned} I(S;M) &\leq I(S;T,P,A) \\ &\quad \text{(Cover and Thomas, 2006, Thm. 2.8.1)} \\ &\leq I(S;T) + I(S;P) + I(S;A) \\ &\quad \text{(Duchi, 2019, Inequality 2.1.7)} \end{aligned}$$

□

	News					Scientific		Social Media	Meeting	Script
	CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
CMP$_w$	0.056	0.426	0.122	0.080	0.092	0.151	0.062	0.113	0.026	0.011
CMP$_s$	0.107	0.116	0.129	0.028	0.096	0.170	0.067	0.161	0.020	0.008
TS	0.160	0.187	0.197	0.183	0.194	0.151	0.151	0.177	0.213	0.195
ABS$_1$	0.074	0.148	0.183	0.174	0.146	0.116	0.055	0.170	0.060	0.064
RED	0.046	-	0.068	-	-	0.036	0.031	0.090	0.037	0.044
SC	0.124	-	0.116	-	-	0.037	0.042	0.172	0.056	0.075

Table 9: Aspect-level standard deviations for each dataset. Redundancy and semantic coherence are not reported for datasets with $> 95\%$ single-sentence summaries.

E Additional Statistics

In the main paper, we report the average score for each metric on each dataset. To complement reporting the mean, we report the standard deviation for each metric on each dataset in Table 9.