SummEval: Re-evaluating Summarization Evaluation

Alexander R. Fabbri[†*]   **Wojciech Kryściński**[‡*]
Bryan McCann[‡]   **Caiming Xiong**[‡]   **Richard Socher**[‡]   **Dragomir Radev**[†‡]

[†]Yale University   [‡]Salesforce Research
{alexander.fabbri,dragomir.radev}@yale.edu,
{kryscinski,cxiong}@salesforce.com
richard@socher.org
bryan.mccann.is@gmail.com

Abstract

The scarcity of comprehensive up-to-date studies on evaluation metrics for text summarization and the lack of consensus regarding evaluation protocols continue to inhibit progress. We address the existing shortcomings of summarization evaluation methods along five dimensions: 1) we re-evaluate 14 automatic evaluation metrics in a comprehensive and consistent fashion using neural summarization model outputs along with expert and crowd-sourced human annotations; 2) we consistently benchmark 23 recent summarization models using the aforementioned automatic evaluation metrics; 3) we assemble the largest collection of summaries generated by models trained on the CNN/DailyMail news dataset and share it in a unified format; 4) we implement and share a toolkit that provides an extensible and unified API for evaluating summarization models across a broad range of automatic metrics; and 5) we assemble and share the largest and most diverse, in terms of model types, collection of human judgments of model-generated summaries on the CNN/Daily Mail dataset annotated by both expert judges and crowd-source workers. We hope that this work will help promote a more complete evaluation protocol for text summarization as well as advance research in developing evaluation metrics that better correlate with human judgments.

1   Introduction

Text summarization aims to compress long document(s) into a short, fluent, and human-readable form that preserves the most salient information from the source document.

*Equal contributions from authors

The field has benefited from advances in neural network architectures (Sutskever et al., 2014; Bahdanau et al., 2014; Vinyals et al., 2015; Vaswani et al., 2017) as well as the availability of large-scale datasets (Sandhaus, 2008; Hermann et al., 2015; Grusky et al., 2018; Narayan et al., 2018). Recent advances in pretrained language models, such as BERT (Devlin et al., 2019), have motivated a corresponding shift to pretraining methods in summarization (Liu and Lapata, 2019; Zhang et al., 2019b; Dong et al., 2019; Ziegler et al., 2019; Raffel et al., 2019; Lewis et al., 2019).

A standard dataset for training summarization models is the CNN/DailyMail corpus (Hermann et al., 2015), originally a question answering task, which was repurposed for summarization by Nallapati et al. (2016). The dataset consists of news articles and associated human-created bullet-point summaries. The ROUGE (Lin, 2004b) metric, which measures lexical overlap between generated and target summaries, is then typically used together with crowd-sourced human annotations for model evaluation. While the current setup has become standardized, we believe several factors prevent a more complete comparison of models, thus negatively impacting the progress of the field.

As noted by Hardy et al. (2019), recent papers vastly differ in their evaluation protocol. Existing work often limits model comparisons to only a few baselines and offers human evaluations which are largely inconsistent with prior work. Additionally, despite problems associated with ROUGE when used outside of its original setting (Liu and Liu, 2008; Cohan and Goharian, 2016) as well as the introduction of many variations on ROUGE (Zhou et al., 2006; Ng and Abrecht, 2015; Ganesan, 2015; ShafieiBavani et al., 2018) and other text generation metrics (Peyrard, 2019; Zhao et al., 2019; Zhang et al., 2020; Scialom

391

et al., 2019; Clark et al., 2019), ROUGE has remained the default automatic evaluation metric. We believe that the shortcomings of the current evaluation protocol are partially caused by the lack of easy-to-use resources for evaluation, both in the form of simplified evaluation toolkits and large collections of model outputs.

In parallel, there is an issue with how evaluation metrics are evaluated themselves. Many of the currently used metrics were developed and assessed using the Document Understanding Conference (DUC) and Text Analysis Conference (TAC) shared-tasks datasets (Dang and Owczarzak 2008, 2009). However, it has recently been shown that the mentioned datasets contain human judgments for model outputs scoring on a lower scale compared to current summarization systems putting into question the true performance of those metrics in the new setting (Peyrard, 2019).

We address these gaps in complementary ways: 1) We re-evaluate 14 automatic evaluation metrics in a comprehensive and consistent fashion using outputs from recent neural summarization models along with expert and crowd-sourced human annotations; 2) We consistently benchmark 23 recent summarization models using the aforementioned automatic evaluation metrics; 3) We release aligned summarization model outputs from 23 papers (44 model outputs) published between 2017 and 2019 trained on the CNN/DailyMail dataset to allow for large-scale comparisons of recent summarization models; 4) We release a toolkit of 14 evaluation metrics with an extensible and unified API to promote the reporting of additional metrics in papers; 5) We collect and release expert, as well as crowd-sourced, human judgments for 16 model outputs on 100 articles over 4 dimensions to further research into human-correlated evaluation metrics. Code and data associated with this work is available at `https://github.com/Yale-LILY/SummEval`.

2   Related Work

Previous work examining the research setup of text summarization can be broadly categorized into three groups, based on the subject of analysis: evaluation metrics, datasets, and models.

Dealing with evaluation methods, Lin (2004a) examined the effectiveness of the ROUGE metric in various DUC tasks. The authors concluded that evaluating against multiple references results in higher correlation scores with human judgments —however, a single-reference setting is sufficient for the metric to be effective. Owczarzak et al. (2012) studied the effects of inconsistencies in human annotations on the rankings of evaluated summarization systems. Results showed that system-level rankings were robust against annotation inconsistencies, but summary-level rankings were not stable in such settings and largely benefit from improving annotator consistency. Rankel et al. (2013) analyzed the performance of different variants of the ROUGE metric using TAC datasets. The authors found that higher-order and less commonly reported ROUGE settings showed a higher correlation with human judgments. In a similar line of work, Graham (2015) conducted a large-scale study of the effectiveness of different ROUGE metric variants and compared it against the BLEU metric on the DUC datasets. Its results highlighted several superior, non-standard ROUGE settings that achieved strong correlations with human judgments on model-generated summaries. In Chaganty et al. (2018), the authors investigated using an automatic metric to reduce the cost of human evaluation without introducing bias. Together with the study, the authors released a set of human judgments over several model outputs, limited to a small set of model types. Peyrard (2019) showed that standard metrics are in agreement when dealing with summaries in the scoring range found in TAC summaries, but vastly differ in the higher-scoring range found in current models. The authors reported that additional human annotations on modern model outputs are necessary to conduct a conclusive study of evaluation metrics. Hardy et al. (2019) underscore the differences in approaches to human summary evaluation while proposing a highlight-based reference-less evaluation metric. Other work has examined the problems with applying ROUGE in settings such as meeting summarization (Liu and Liu, 2008) and summarization of scientific articles (Cohan and Goharian, 2016). We build upon this line of research by examining the performance of several automatic evaluation methods, including ROUGE and its variants, against the performance of expert human annotators.

In relation to datasets, Dernoncourt et al. (2018) presented a detailed taxonomy of existing summarization datasets. The authors highlighted the differences in formats of available corpora and

called for creating a unified data standard. In a similar line of research, Grusky et al. (2018) offered a thorough analysis of existing corpora, focusing their efforts on news summarization datasets. The authors also introduced several metrics for evaluating the extractiveness of summaries that are included in the toolkit implemented as part of this work. Kryściński et al. (2020) showed that news-related summarization datasets, such as CNN/DailyMail, contain strong layout biases. The authors revealed that datasets in the current format, where each news article is associated with a single reference summary, leave the task of summarization underconstrained. The paper also highlighted the problem of noisy, low-quality data in automatically collected news datasets.

Looking into models, Zhang et al. (2018a) analyzed the level of abstraction of several recent abstractive summarization models. The authors showed that word-level extractive models achieved a similar level of abstraction to fully abstractive models. In Kedzie et al. (2018), the authors examined the influence of various model components on the quality of content selection. The study revealed that in the current setting the training signal is dominated by biases present in summarization datasets preventing models from learning accurate content selection. Kryściński et al. (2020) investigate the problem of factual correctness of text summarization models. The authors concluded that the issue of hallucinating facts touches up to 30% of generated summaries and list common types of errors made by generative models. Closely related to that work, Maynez et al. (2020) conducted a large-scale study of abstractive summarizers from the perspective of faithfulness. The authors reached similar conclusions, stating that improving factual faithfulness is a critical issue in summarization. The results also showed that currently available evaluation methods, such as ROUGE and BertScore, are not sufficient to study the problem at hand. Durmus et al. (2020) and Wang et al. (2020) similarly examine faithfulness evaluation, both proposing question answering frameworks as a means of evaluating factual consistency.

Insights and contributions coming from our work are complementary to the conclusions of previous efforts described in this section. To the best of our knowledge, this is the first work in neural text summarization to offer a large-scale,

consistent, side-by-side re-evaluation of summarization model outputs and evaluation methods. We also share resources that we hope will prove useful for future work in analyzing and improving summarization models and metrics.

Shortly before publishing this paper, a library for developing summarization metrics was released by Deutsch and Roth (2020). Our toolkit is complementary to their work as their toolkit includes only 3 of our 12 evaluation metrics.

3 Evaluation Metrics and Summarization Models

We briefly introduce metrics included in our evaluation toolkit as well as the summarization models for which outputs were collected at the time of releasing this manuscript.

3.1 Evaluation Metrics

Our selection of evaluation methods includes several recently introduced metrics that have been applied to both text generation and summarization, standard machine translation metrics, and other miscellaneous performance statistics.

ROUGE (Lin, 2004b), (Recall-Oriented Understudy for Gisting Evaluation), measures the number of overlapping textual units (n-grams, word sequences) between the generated summary and a set of gold reference summaries.

ROUGE-WE (Ng and Abrecht, 2015) extends ROUGE by using soft lexical matching based on the cosine similarity of Word2Vec (Mikolov et al., 2013) embeddings.

S^3 (Peyrard et al., 2017) is a model-based metric that uses previously proposed evaluation metrics, such as ROUGE, JS-divergence, and ROUGE-WE, as input features for predicting the evaluation score. The model is trained on human judgment datasets from TAC conferences.

BertScore (Zhang et al., 2020) computes similarity scores by aligning generated and reference summaries on a token-level. Token alignments are computed greedily to maximize the cosine similarity between contextualized token embeddings from BERT.

MoverScore (Zhao et al., 2019) measures the semantic distance between a summary and reference text by making use of the Word Mover's Distance (Kusner et al., 2015) operating over n-gram embeddings pooled from BERT representations.

Sentence Mover's Similarity (SMS) (Clark et al., 2019) extends Word Mover's Distance to view documents as a bag of sentence embeddings as well as a variation which represents documents as both a bag of sentences and a bag of words.

SummaQA (Scialom et al., 2019) applies a BERT-based question-answering model to answer cloze-style questions using generated summaries. Questions are generated by masking named entities in source documents associated with evaluated summaries. The metric reports both the F1 overlap score and QA-model confidence.

BLANC (Vasilyev et al., 2020) is a reference-less metric that measures the performance gains of a pre-trained language model given access to a document summary while carrying out language understanding tasks on the source document's text.

SUPERT (Gao et al., 2020) is a reference-less metric, originally designed for multi-document summarization, which measures the semantic similarity of model outputs with pseudo-reference summaries created by extracting salient sentences from the source documents, using soft token alignment techniques.

BLEU (Papineni et al., 2002) is a corpus-level precision-focused metric that calculates n-gram overlap between a candidate and reference utterance and includes a brevity penalty. It is the primary evaluation metric for machine translation.

CHRF (Popović, 2015) calculates character-based n-gram overlap between model outputs and reference documents.

METEOR (Lavie and Agarwal, 2007) computes an alignment between candidate and reference sentences by mapping unigrams in the generated summary to 0 or 1 unigrams in the reference, based on stemming, synonyms, and paraphrastic matches. Precision and recall are computed and reported as a harmonic mean.

CIDEr (Vedantam et al., 2015) computes {1–4}-gram co-occurrences between the candidate and reference texts, down-weighting common n-grams and calculating cosine similarity between the n-grams of the candidate and reference texts.

Data Statistics: Grusky et al. (2018) define three measures of the extractiveness of a dataset. *Extractive fragment coverage* is the percentage of words in the summary that are from the source article, measuring the extent to which a summary is a derivative of a text. *Density* is defined as the average length of the extractive fragment to which each summary word belongs. *Compression ratio* is defined as the word ratio between the articles and its summaries: In addition to these measures, we also include the percentage of n-grams in the summary not found in the input document as a *novelty* score and the percentage of n-grams in the summary which repeat as a score of *redundancy*. For a comprehensive explanation of each metric, please refer to the corresponding paper.

3.2 Summarization Models

We broadly categorize the models included in this study into extractive and abstractive approaches. For each model, we provide a model code (M*) as well as a descriptive model name, which will allow for easy matching with the released data.

Extractive Methods

M1 - **NEUSUM** (Zhou et al., 2018) jointly scores and selects sentences by first building a hierarchical representation of a document and considering the partially outputted summary at each time step.

M2 - **BanditSum** (Dong et al., 2018) treats extractive summarization as a contextual bandit problem where the document is the context and the sequence of sentences to include in the summary is the action.

M3 - **LATENT** Zhang et al. (2018b) propose a latent variable extractive model which views rele-vance labels of sentences in a document as binarylatent variables.

M4 - **REFRESH** Narayan et al. (2018) propose using REINFORCE (Williams, 1992) to extract summaries, approximating the search space during training by limiting to combinations of individually high-scoring sentences.

M5 - **RNES** Wu and Hu (2018) propose a coherence model to capture cross-sentence coherence, combining output from the coherence model and ROUGE scores as a reward in a REINFORCE framework.

M6 - **JECS** (Xu and Durrett, 2019) first extracts sentences from a document and then scores possible constituency-based compressed units to produce the final compressed summary.

M7 - **STRASS** (Bouscarrat et al., 2019) extracts a summary by selecting the sentences with the closest embeddings to the document embedding, learning a transformation to maximize the similarity between the summary and the ground truth reference.

Abstractive Methods

M8 - **Pointer Generator** See et al. (2017) propose a variation of encoder-decoder models, the Pointer Generator Network, where the decoder can choose to generate a word from the vocabulary or copy a word from the input. A coverage mechanism is also proposed to prevent repeatedly attending to the same part of the source document.

M9 - **Fast-abs-rl** Chen and Bansal (2018) propose a model which first extracts salient sentences with a Pointer Network and rewrites these sentences with a Pointer Generator Network. In addition to maximum likelihood training, a ROUGE-L reward is used to update the extractor via REINFORCE (Williams, 1992).

M10 - **Bottom-Up** Gehrmann et al. (2018) introduce a bottom–up approach whereby a content selection model restricts the copy attention distribution of a pretrained Pointer Generator Network during inference.

M11 - **Improve-abs** Kryściński et al. (2018) extend the model of Paulus et al. (2017) by augmenting the decoder with an external LSTM language model and add a novelty RL-based objective during training.

M12 - **Unified-ext-abs** Hsu et al. (2018) propose to use the probability output of an extractive model as sentence-level attention to modify word-level attention scores of an abstractive model, introducing an inconsistency loss to encourage consistency between these two levels of attention.

M13 - **ROUGESal** Pasunuru and Bansal (2018) propose a keyphrase-based salience reward as well as an entailment-based reward in addition to using a ROUGE-based reward in a REINFORCE setting, optimizing rewards simultaneously in alternate mini-batches.

M14 - **Multi-task (Ent + QG)** Guo et al. (2018) propose question generation and entailment generation as auxiliary tasks in a multi-task framework along with a corresponding multi-task architecture.

M15 - **Closed book decoder** Jiang and Bansal (2018) build upon a Pointer Generator Network by adding copy-less and attention-less decoder during training time to force the encoder to be more selective in encoding salient content.

M16 - **SENECA** Sharma et al. (2019) propose to use entity-aware content selection module and an abstractive generation module to generate the final summary.

M17 - **T5** Raffel et al. (2019) perform a systematic study of transfer learning techniques and apply their insights to a set of tasks all framed as text-input to text-output generation tasks, including summarization.

M18 - **NeuralTD** Böhm et al. (2019) learn a reward function from 2,500 human judgments that is used in a reinforcement learning setting.

M19 - **BertSum-abs** Liu and Lapata (2019) introduce a novel document-level encoder on top of BERT (Devlin et al., 2019), over which they introduce both an extractive and an abstractive model.

M20 - **GPT-2** Ziegler et al. (2019) build off of GPT-2 (Radford et al., 2019) and fine-tune the model by using human labels of which of four sampled summaries is the best to direct fine-tuning in a reinforcement learning framework.

M21 - **UniLM** Dong et al. (2019) introduce a model pretrained on three language modeling tasks: unidirectional, bidirectional, and sequence-to-sequence prediction. It is thus applicable to natural language understanding tasks and generation tasks such as abstractive summarization.

M22 - **BART** Lewis et al. (2019) introduce a denoising autoencoder for pretraining sequence to sequence tasks which is applicable to both natural language understanding and generation tasks.

M23 - **Pegasus** Zhang et al. (2019a) introduce a model pretrained with a novel objective function designed for summarization by which important sentences are removed from an input document and then generated from the remaining sentences.

4 Resources

We now describe the resources collected and released together with this manuscript.

4.1 Model Outputs

The model output collection contains summaries associated with 23 recent papers on neural text summarization described in Section 3.2. We obtained a total of 44 model outputs, as many papers include variations of the main model. All models were trained on the CNN/DailyMail news corpus and the collected summaries were generated using the test split of the dataset without constraints limiting the output length. Outputs were solicited from the authors of papers to ensure comparability between results presented in this

paper with those in the original works. They are shared publicly with the consent of the authors.

Model outputs were transformed into a unified format and are shared with IDs of the original CNN/DailyMail examples so that generated summaries can be matched with corresponding source articles. Pairing model outputs with original articles was done using a heuristic approach that relied on aligning reference summaries. The pairing process revealed that 38 examples in the CNN/DailyMail test split contained duplicate reference summaries preventing those examples to be correctly aligned. However, this problem involves only 0.3% of the available data and should not have a significant impact on downstream results. IDs of duplicate examples are provided together with the data.

4.2 Evaluation Toolkit

The evaluation toolkit contains 14 automatic evaluation metrics described in Section 3.1 consolidated into a Python package. The package provides a high-level, easy-to-use interface unifying all of the underlying metrics. For each metric, we implement both `evaluate_example` and `evaluate_batch` functions that return the metric's score on example- and corpus-levels accordingly. Function inputs and outputs are also unified across all metrics to streamline multi-metric evaluation and result processing. The toolkit comes with a standard configuration resembling the most popular settings for each of the metrics to enable easy, out-of-the-box use. However, each metric can be further configured using external `gin` configuration files. We also provide a command-line tool to evaluate a summarization model with several metrics in parallel.

4.3 Human Annotations

The collection of human annotations contains summary evaluations of 16 recent neural summarization models solicited from crowd-sourced and expert judges. Annotations were collected for 100 articles randomly picked from the CNN/DailyMail test set. To ensure high quality of annotations, each summary was scored by 5 crowd-sourced and 3 expert workers, amounting to 12800 summary-level annotations. Model outputs were evaluated along the following four dimensions, as in Kryściński et al. (2019):

Coherence - the collective quality of all sentences. We align this dimension with the DUC quality question (Dang, 2005) of structure and coherence whereby ''the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.''

Consistency - the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts.

Fluency - the quality of individual sentences. Drawing again from the DUC quality guidelines, sentences in the summary ''should have no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.''

Relevance - selection of important content from the source. The summary should include only important information from the source document. Annotators were instructed to penalize summaries that contained redundancies and excess information.

The data collection interface provided judges with the source article and associated summaries grouped in sets of 5. Each group of summaries contained the reference summary associated with the source article to establish a common point of reference between groups. Summary grouping and order within groups were randomized for each annotator. Judges were asked to rate the summaries on a Likert scale from 1 to 5 (higher better) along the four mentioned dimensions.

Crowd-sourced annotators were hired through the Amazon Mechanical Turk platform. The hiring criteria were set to a minimum of 10000 approved HITs and an approval rate of 97% or higher. Geographic constraints for workers were set to United States, United Kingdom, and Australia to ensure that summaries were evaluated by native English speakers. Compensation was carefully calculated to ensure an average wage of 12 USD per hour.

Gillick and Liu (2010) showed that summary judgments obtained through non-experts may differ greatly from expert annotations and could exhibit worse inter-annotator agreement. As a result, in addition to the hired crowd-sourced workers, we enlisted three expert annotators who have written papers on summarization either for academic conferences (2) or as part of a senior

396

thesis (1). The expert annotators were asked to evaluate the same set of summaries under the same instructions as the hired crowd-sourced workers. For expert judgments, we proceeded with two rounds of annotation to correct any obvious mistakes as well as to confirm judgments and ensure a higher quality of annotations. In the second round, annotators were asked to check all examples for which their score of a dimension differed from another annotator by more than 2 points and where the other annotators were within 1 point of each other. In cases where a score differed by more than 2 points for which such a pattern did not exist, all annotators examined the annotation. When re-evaluating examples, judges were allowed to see scores assigned by other expert annotators in the first round of annotations. While such a setting could undermine the wisdom of the crowd and shift the re-assigned scores towards the average judgment from the first round, we encouraged experts to remain critical and discuss contested examples when necessary. For completeness, the data collection user interface and additional details regarding the data collection process are presented in the Appendix.

5 Metric Re-evaluation

5.1 Human Annotations

Considering the concerns raised in previous work (Gillick and Liu, 2010) about the quality differences between crowd-sourced and expert annotations we study this issue using the human annotations collected as part of this work.

To evaluate the inter-annotator agreement of collected crowd-sourced and expert annotations we computed the Krippendorff's alpha coefficient (Krippendorff, 2011). We found the inter-annotator interval kappa to be below an acceptable range—0.4920 and 0.4132 for the crowd-sourced workers and the first round of expert annotations, respectively. However, the second round of expert annotations improved the inter-annotator agreement, achieving a kappa coefficient of 0.7127. For further insights, we computed standard deviations of annotator scores within the respective groups and present histograms of those statistics in Figure 1. Plots of crowd-sourced annotations show strong similarities across all evaluated dimensions. Such an effect could be caused by an insufficient distinction made by the annotators between the 4

scored axes, where the overall quality of a summary biased scores of the individual dimensions. The histograms also show that while the second round of expert annotations lowered the standard deviation of scores and substantially increased inter-annotator agreement, relevance and coherence remained the most disagreed on dimensions between experts. This could be attributed to the subjective nature of relevance and coherence as an evaluation dimensions (Kryściński et al., 2020).

To assess the similarity of annotations between the crowd-sourced and expert annotators, we averaged the assigned scores per example within the respective annotator groups and computed Pearson's correlation coefficient. The statistic returned a value close to 0, indicating no correlation between expert and crowd-sourced judges.

We also manually inspected the human annotations and present examples of annotated summaries, both generated and reference, as well as the differences in human judgments in Table 1a. The first row shows a well written, comprehensive summary. The high quality of the summary is reflected by top scores assigned by expert annotators, while being rated as average by crowd-sourced workers. The second row shows a summary with ambiguous pronoun usage and factual inconsistencies. The errors result in a decrease in coherence, consistency, and relevance scores in the expert annotations, but do not see a corresponding decrease in crowd-worker annotations. The third row presents a factually correct summary that contains token and phrase repetitions. The errors were caught by the expert annotators resulting in a low fluency score, while crowd-sourced annotators incorrectly classified them as issues with factual consistency. These examples again illustrate the disparities in the understanding of evaluated dimensions between judges and underscore our observation above about the uniformity of crowd-sourced annotations; the crowd-sourced annotations tend to be similar across quality dimensions even when distinctions exist, which are captured in the expert annotations.

Results presented in this section highlight the difficulties of crowd-sourcing high-quality annotations and the necessity for protocols for improving human evaluation in text summarization.

5.2 Automatic Metrics

Many automatic metrics have been proposed for evaluating both summarization and other text

Generated Summaries	Expert scores (avg.)	Crowd-worker scores (avg.)
the queen's guard was left red-faced after he slipped on a he manhole cover lost his footing and slid sideways, knocking his bearskin on the side . the embarrassed soldier quickly scrambled to his feet as his colleagues marched past as if nothing had happened . tourist david meadwell recorded the unscheduled manouevre outside buckingham palace on thursday afternoon .	Coh: 5.0 Con: 5.0 Flu: 5.0 Rel: 5.0	Coh: 3.4 Con: 3.8 Flu: 3.4 Rel: 3.8
holidaymaker david meadwell recorded the unscheduled manouevre outside buckingham palace . he lost his footing and slid sideways, knocking bearskin on the side of the box . queen 's guard was left red-faced after he slipped on manhole cover . the entire incident was caught on a manhole cover . the embarrassed soldier quickly scrambled to his feet as his colleagues marched past .	Coh: 2.7 Con: 2.0 Flu: 4.7 Rel: 3.7	Coh: 3.2 Con: 3.4 Flu: 3.4 Rel: 4.0
buckingham palace guard slipped on manhole cover in front of hundreds of horrified tourists. the queen 's guard was left red-faced after he slipped on a manhole cover . he lost his footing and dropped his rifle on the side of the box and dropping his rifle . the incident was caught on camera camera camera . the guard is thought to have slipped because of metal shutters nailed to the soles of his boots .	Coh: 3.3 Con: 5.0 Flu: 1.7 Rel: 4.3	Coh: 3.0 Con: 3.2 Flu: 2.8 Rel: 3.2

(a) Generated summary examples illustrate common problems found in model outputs, such as ambiguous pronouns, incorrect references, and repetitive content.

Reference Summaries	Expert scores (avg.)	Crowd-worker scores (avg.)
river plate admit they ' dream ' of manchester united striker radamel falcao . the colombia international spent eight years with the argentine club . falcao has managed just four goals in 19 premier league appearances . read : falcao still ' has faith ' that he could continue at man utd next season . click here for the latest manchester united news.	Coh: 3.0 Con: 2.0 Flu: 5.0 Rel: 2.3	Coh: 3.0 Con: 3.6 Flu: 3.0 Rel: 4.4
the incident occurred on april 7 north of poland in the baltic sea . u.s. says plane was in international airspace . russia says it had transponder turned off and was flying toward russia	Coh: 2.0 Con: 1.7 Flu: 3.0 Rel: 2.3	Coh: 4.0 Con: 3.4 Flu: 4.2 Rel: 3.6

(b) Reference summaries highlight issues found in theCNN/DailyMail dataset, such as click-baits and references to other articles as well as unreferenced dates and lowcoherence caused by concatenating bullet-point summaries.

Table 1: Example summaries with the corresponding averaged expert and crowd-sourced annotations for *coherence*, *consistency*, *fluency*, and *relevance*. Expert annotations better differentiate coherence, consistency, and fluency among the examples when compared to the crowd-sourced annotations.

generation models. However, the field lacks a comprehensive study that would offer a consistent side-by-side comparison of their performance. We address this issue with the following experiments.

In Table 2 we show Kendall's tau rank correlations between automatic metrics and human judgments calculated on a system-level following Louis and Nenkova (2013). The statistics were computed using the available expert annotations to avoid possible quality problems associated with crowd-sourced ratings, as highlighted in the previous subsection. Automatic metrics were computed
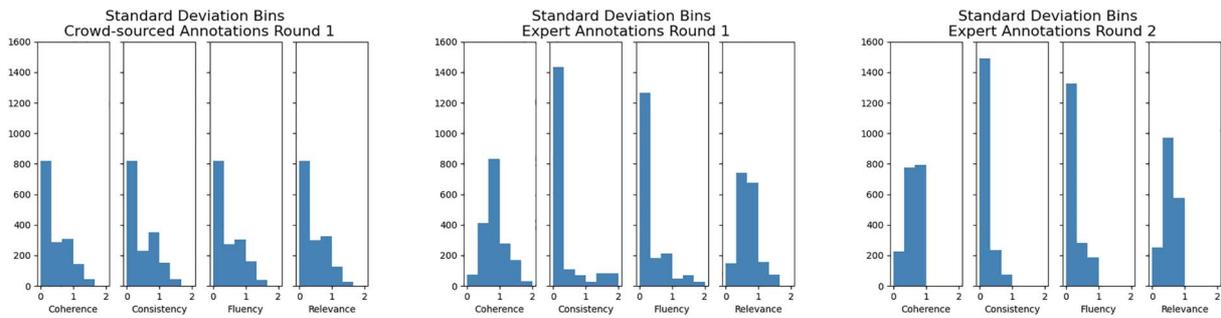
Figure 1: Histogram of standard deviations of inter-annotator scores between: crowd-sourced annotations, first round expert annotations, and second round expert annotations, respectively.
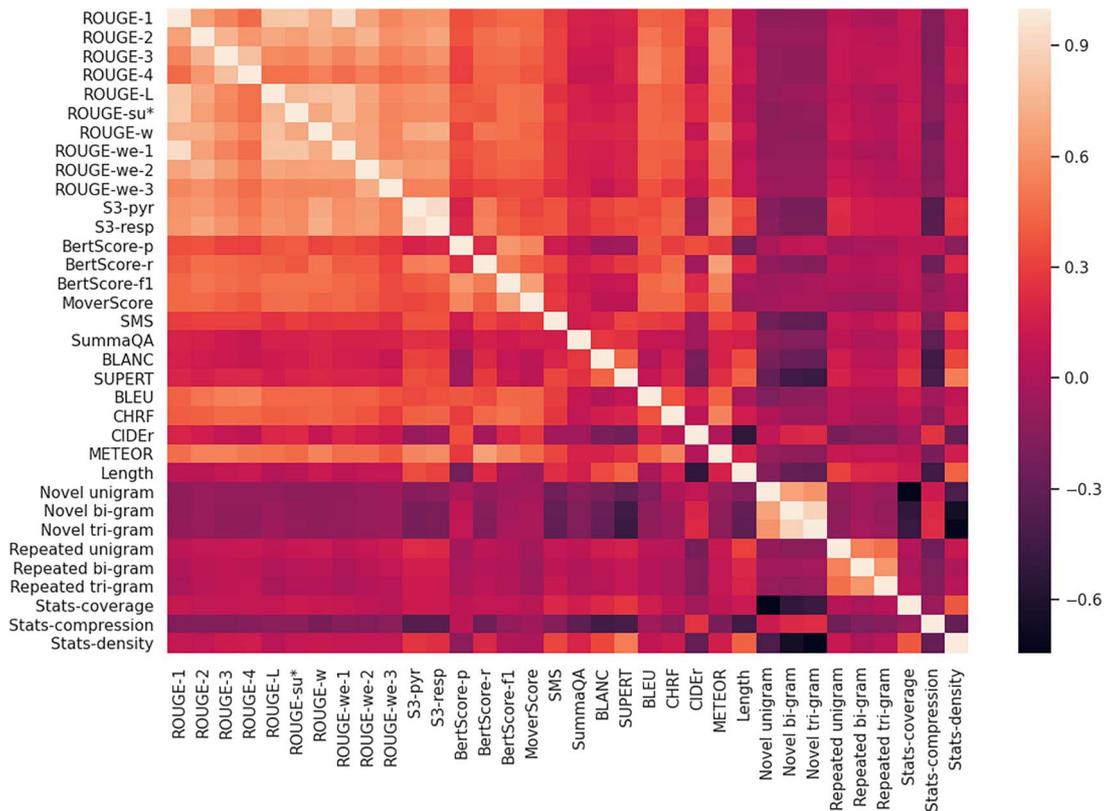

Figure 2: Pairwise Kendall's tau correlations for all automatic evaluation metrics.

in a multi-reference setting, using the original reference summary included in the CNN/DailyMail dataset and 10 additional summaries coming from Kryściński et al. (2020), and the length of model outputs was not constrained. We report correlations without differentiating between abstractive and extractive models, as most metrics did not exhibit large differences in correlation when reported separately.

Correlation results show several trends. We find that most metrics have the lowest correlation within the coherence dimension, where the correlation strength can be classified as weak or moderate. This finding follows intuition as the

majority of metrics rely on hard or soft subsequence alignments, which do not measure well the interdependence between consecutive sentences. Low and moderate correlation scores were also found for the relevance dimension. As discussed in the previous subsection, such trends could result from the inherent subjectiveness of the dimension and the difficulty of collecting consistent human annotations. Model correlations increase considerably across the consistency and fluency dimensions. Although unexpected, the strong correlation with consistency could be attributed to the low abstractiveness of most neural models, which could increase the effectiveness of metrics using

Metric	Coherence	Consistency	Fluency	Relevance
ROUGE-1	0.2500	0.5294	**0.5240**	0.4118
ROUGE-2	0.1618	0.5882	0.4797	0.2941
ROUGE-3	0.2206	**0.7059**	**0.5092**	0.3529
ROUGE-4	**0.3088**	0.5882	**0.5535**	0.4118
ROUGE-L	0.0735	0.1471	0.2583	0.2353
ROUGE-su*	0.1912	0.2941	0.4354	0.3235
ROUGE-w	0.0000	0.3971	0.3764	0.1618
ROUGE-we-1	**0.2647**	0.4559	**0.5092**	0.4265
ROUGE-we-2	−0.0147	0.5000	0.3026	0.1176
ROUGE-we-3	0.0294	0.3676	0.3026	0.1912
S^3-pyr	−0.0294	0.5147	0.3173	0.1324
S^3-resp	−0.0147	0.5000	0.3321	0.1471
BertScore-p	0.0588	−0.1912	0.0074	0.1618
BertScore-r	0.1471	**0.6618**	0.4945	0.3088
BertScore-f	0.2059	0.0441	0.2435	**0.4265**
MoverScore	0.1912	−0.0294	0.2583	0.2941
SMS	0.1618	0.5588	0.3616	0.2353
SummaQAˆ	0.1176	**0.6029**	0.4059	0.2206
BLANCˆ	0.0735	0.5588	0.3616	0.2647
SUPERTˆ	0.1029	0.5882	0.4207	0.2353
BLEU	0.1176	0.0735	0.3321	0.2206
CHRF	**0.3971**	0.5294	0.4649	**0.5882**
CIDEr	0.1176	−0.1912	−0.0221	0.1912
METEOR	0.2353	**0.6324**	**0.6126**	**0.4265**
Lengthˆ	−0.0294	0.4265	0.2583	0.1618
Novel unigramˆ	0.1471	−0.2206	−0.1402	0.1029
Novel bi-gramˆ	0.0294	−0.5441	−0.3469	−0.1029
Novel tri-gramˆ	0.0294	−0.5735	−0.3469	−0.1324
Repeated unigramˆ	−0.3824	0.1029	−0.0664	−0.3676
Repeated bi-gramˆ	−0.3824	−0.0147	−0.2435	−0.4559
Repeated tri-gramˆ	−0.2206	0.1471	−0.0221	−0.2647
Stats-coverageˆ	−0.1324	0.3529	0.1550	−0.0294
Stats-compressionˆ	0.1176	−0.4265	−0.2288	−0.0147
Stats-densityˆ	0.1618	**0.6471**	0.3911	0.2941

Table 2: Kendall's tau correlation coefficients of expert annotations computed on a system-level along four quality dimensions with automatic metrics using 11 reference summaries per example. ˆ denotes metrics which use the source document. The five most-correlated metrics in each column are bolded.

higher-order n-gram overlap, such as ROUGE-3 or Extractive Density. Referring back to the previous subsection, both of the mentioned dimensions achieved high inter-annotator agreement between expert judges which could also positively affect the correlation scores. Additionally, the results show a substantially higher correlation between all evaluated dimensions and ROUGE scores computed for higher-order n-grams in comparison to ROUGE-L, which corroborates with findings of Rankel et al. (2013).

To examine the dependencies between different metrics, we computed Kendall's tau rank correlation coefficients, pairwise, between all metrics. Results are presented as a correlation matrix in Figure 2. Following intuition, we observe a

strong correlation between all metrics that compute, implicitly or explicitly, the lexical overlap between generated and reference summaries. Metrics measuring the n-gram novelty and repetitiveness show a weak negative correlation with all ROUGE-related metrics. Length as a feature is weakly correlated with most metrics apart from S^3, BLANC, and SuPERT, which might suggest the mentioned metrics favor longer summaries. Worth noting is also the weak correlation of reference-less SummaQA, BLANC, and SuPERT metrics with most other evaluated metrics.

Results presented in this section highlight the evaluation dimensions that are not reliably covered by currently available metrics and pave the way for future work in model evaluation.

6 Model Re-evaluation

We now turn to an analysis of model scores across human evaluations and automatic metrics. The evaluated models were released between 2017 and 2019, represent different approaches to summarization: abstractive, extractive, and hybrid, and their architectures reflect the trends in summarization research. Although in many cases we obtained multiple variants of the same model, in the study we focus on the versions with the highest ROUGE-L scores.

Table 3 contains the results of human evaluation across the four dimensions described in Section 4.3. Scores for ground truth summaries are included as a point of reference. We find that pretrained models such as Pegasus, BART, and T5 consistently performed best on most dimensions. Notably, the mentioned models scored highest on consistency and fluency while obtaining lower scores for relevance and coherence. Scores for extractive models highlight the known shortcomings of such approaches, which are lack of coherence of summaries and issues with selecting relevant content. Abstractive model ratings show an increasing trend with respect to the date of publication. This is a promising result as it suggests that the quality of models is improving with time. Worth noting is also the fact that reference summaries did not score well on consistency, coherence, and relevance. Upon examination of the annotations, we found that the reference summaries often contained extraneous information, such as hyperlinks and click-bait descriptions of other articles. As this information was not present

Method	Coherence	Consistency	Fluency	Relevance
CNN/DM Reference Summary	3.26	4.47	4.79	3.77
Extractive Models				
M0 - **LEAD-3**	**4.16**	4.98	4.94	4.14
M1 - **NEUSUM**	3.22	4.98	4.90	3.82
M2 - **BanditSum**	3.28	4.99	4.83	3.81
M5 - **RNES**	3.71	**4.97**	4.81	4.06
Abstractive Models				
M8 - **Pointer Generator**	3.29	4.65	4.79	3.55
M9 - **Fast-abs-rl**	2.38	4.67	4.50	3.52
M10 - **Bottom-Up**	2.73	4.25	4.42	3.38
M11 - **Improve-abs**	2.28	3.27	3.65	3.15
M12 - **Unified-ext-abs**	3.60	**4.96**	4.85	3.85
M13 - **ROUGESal**	3.44	4.82	4.86	3.83
M14 - **Multi-task (Ent + QG)**	3.20	4.90	4.74	3.63
M15 - **Closed book decoder**	3.35	**4.95**	4.80	3.67
M17 - **T5**	4.00	4.93	**4.93**	4.23
M20 - **GPT-2** (zero shot)[1]	3.63	3.40	3.97	3.30
M22 - **BART**	**4.18**	4.94	**4.90**	**4.25**
M23 - **Pegasus** (C4)	**4.16**	4.91	**4.88**	**4.26**
M23 - **Pegasus** (dynamic mix)	4.09	4.85	4.79	**4.27**

Table 3: Human ratings of summaries along four evaluation dimensions, averaged over three expert annotators, broken down by extractive and abstractive models. The M* codes follow the notation described in Section 3.2. The three highest-rated models in each column are in bold.

in the source documents nor relevant for the summaries, the annotators interpreted it as hallucinations and assigned lower consistency and relevance scores. Additionally, many reference summaries in the CNN/DailyMail dataset were constructed by naively concatenating bullet-point summaries into contiguous sequences. Such processing steps negatively affected the coherence of examples. Similar trends in human studies of reference summaries were reported by Stiennon et al. (2020). Examples of noisy reference summaries are shown in Table 1b. Table 4 shows scores for model outputs across all automatic evaluation metrics. Parameters of metrics used in this study can be found in the evaluation toolkit repository listed in Section 1. The results align with insights coming from the human evaluation of models. We found that for most metrics, the highest scores were assigned to large models pretrained on vast quantities of data. However, several metrics, such as S^3, SummaQA, SMS, CHRF, and METEOR tended to favor extractive models, assigning the highest scores to their outputs.

Presented results provide a comprehensive perspective on the current state of the field and highlight directions for future modeling work.

7   Conclusions

We introduced SummEval, a set of resources for summarization model and evaluation research that include: a collection of summaries generated by recent summarization models on the CNN/DailyMail dataset, an extensible and unified toolkit for summarization model evaluation, and a diverse collection of human annotations of model outputs collected from the crowd-source and expert annotators. Using the accumulated resources we re-evaluated a broad selection of current models and evaluation metrics in a consistent and comprehensive manner. We hope that this work will prove to be a valuable resource for future research on text summarization evaluation and models. We also encourage the research community to join our efforts by contributing model outputs and extending the evaluation toolkit with new metrics.

---

[1]The zero-shot model was used for evaluation.

Table (a): Model scores from summarization-specific evaluation metrics.

Method	ROUGE-1/2/3/4/L/su*/w	ROUGE-WE-(1/2/3)	S^3 (pyr/resp)	BertScore	MoverScore	SummaQA	SMS	BLANC	SUPERT
Extractive Models									
M0 - LEAD-3	0.3994 / 0.1746 / 0.0990 / 0.0647 / 0.3606 / 0.1377 / 0.2072	0.4049 / 0.2260 / 0.2172	**0.5395** / 0.6328	0.3742	0.1679	**0.1652**	0.1050	0.0480	**0.7259**
M1 - NEUSUM	0.4130 / 0.1893 / 0.1109 / 0.0742 / 0.3768 / 0.1495 / 0.2156	0.4186 / 0.2402 / 0.2310	**0.5562** / **0.6509**	0.3955	0.1839	**0.1700**	**0.1062**	**0.1087**	0.7010
M2 - BanditSum	0.4137 / 0.1868 / 0.1086 / 0.0721 / 0.3759 / 0.1513 / 0.2139	0.4195 / 0.2385 / 0.2300	**0.5339** / 0.6306	0.3938	0.1815	0.1324	**0.1058**	0.0909	**0.7018**
M3 - LATENT	0.4136 / 0.1867 / 0.1085 / 0.0721 / 0.3757 / 0.1512 / 0.2138	0.4194 / 0.2384 / 0.2299	0.5337 / 0.6305	0.3936	0.1814	**0.1645**	0.1058	0.0910	**0.7020**
M4 - REFRESH	0.3972 / 0.1807 / 0.1042 / 0.0690 / 0.3621 / 0.1340 / 0.2129	0.4023 / 0.2318 / 0.2238	**0.6395** / **0.7124**	0.3903	0.1720	0.1088	0.1088	**0.1406**	**0.7526**
M5 - RNES	0.4088 / 0.1878 / 0.1102 / 0.0736 / 0.3719 / 0.1446/ **0.2163**	0.4153 / 0.2395 / 0.2317	**0.6082** / **0.6894**	**0.3997**	0.1802	**0.1794**	**0.1107**	**0.1232**	**0.7434**
M6 - JECS	0.4144 / 0.1846 / 0.1063 / 0.0699 / 0.3760 / 0.1485 / 0.2283	0.4200 / 0.2371 / 0.2283	0.5337 / 0.6284	0.3925	0.1805	0.1644	0.1048	**0.1044**	0.6946
M7 - STRASS	0.3377 / 0.1237 / 0.0650 / 0.0416 / 0.2790 / 0.1052 / 0.1559	0.3477 / 0.1757 / 0.1656	0.3632 / 0.4939	0.3090	0.1079	0.1367	0.1023	0.1042	0.6566
Abstractive Models									
M8 - Pointer Generator	0.3921 / 0.1723 / 0.1003 / 0.0674 / 0.3599 / 0.1435 / 0.1999	0.3990 / 0.2226 / 0.2128	0.4328 / 0.5561	0.3763	0.1643	0.1398	0.0974	0.0704	0.6501
M9 - Fast-abs-rl	0.4057 / 0.1774 / 0.0975 / 0.0616 / 0.3806 / 0.1439 / 0.2112	0.4123 / 0.2302 / 0.2184	0.4818 / 0.5865	0.3918	0.1748	0.1431	0.0847	0.0855	0.6125
M10 - Bottom-Up	0.4124 / 0.1870 / 0.1064 / 0.0695 / **0.3815** / 0.1543 / 0.2084	0.4192 / 0.2400 / 0.2313	0.4450 / 0.5655	0.3964	0.1830	0.1408	0.0925	0.0570	0.6092
M11 - Improve-abs	0.3985 / 0.1720 / 0.0927 / 0.0567 / 0.3730 / 0.1431 / 0.2073	0.4045 / 0.2300 / 0.2228	0.4899 / 0.5897	0.3826	0.1652	0.1341	0.0816	0.0777	0.5972
M12 - Unified-ext-abs	0.4038 / 0.1790 / 0.1039 / 0.0695 / 0.3675 / 0.1484 / 0.2074	0.4097 / 0.2299 / 0.2204	0.4936 / 0.5995	0.3832	0.1739	0.1530	0.1038	0.0962	0.6826
M13 - ROUGESal	0.4016 / 0.1797 / 0.1053 / 0.0709 / 0.3679 / 0.1497 / 0.2058	0.4078 / 0.2294 / 0.2190	0.4643 / 0.5799	0.3837	0.1722	0.1475	0.1009	0.0882	0.6570
M14 - Multi-task (Ent + QG)	0.3952 / 0.1758 / 0.1037 / 0.0705 / 0.3625 / 0.1476 / 0.2007	0.4015 / 0.2253 / 0.2149	0.4246 / 0.5513	0.3759	0.1670	0.1360	0.0982	0.0648	0.6380
M15 - Closed book decoder	0.3976 / 0.1760 / 0.1031 / 0.0696 / 0.3636 / 0.1472 / 0.2033	0.4039 / 0.2263 / 0.2160	0.4591 / 0.5757	0.3783	0.1699	0.1456	0.1009	0.0896	0.6612
M16 - SENECA	0.4151 / 0.1836 / 0.1052 / 0.0681 / 0.3806 / 0.1520 / 0.2112	0.4211 / 0.2369 / 0.2282	0.4735 / 0.5836	0.3907	0.1811	0.1404	0.1005	0.0692	0.6519
M17 - T5	**0.4479** / **0.2205** / **0.1336** / **0.0920** / **0.4172** / **0.1879** / **0.2291**	**0.4543** / **0.2723** / **0.2631**	0.5168 / 0.6294	**0.4450**	**0.2376**	0.1437	0.1046	0.0773	0.6094
M18 - NeuralTD	0.4004 / 0.1762 / 0.1000 / 0.0650 / 0.3723 / 0.1452 / 0.2085	0.4063 / 0.2277 / 0.2187	0.4946 / 0.5975	0.3949	0.1697	0.1440	0.0916	0.0859	0.6290
M19 - BertSum-abs	**0.4163** / **0.1944** / **0.1156** / 0.0785 / 0.3554 / **0.1625** / 0.1979	**0.4230** / **0.2454** / **0.2351**	0.4664 / 0.5855	0.3855	**0.1894**	0.1385	**0.1071**	0.0815	0.6116
M20 - GPT-2 (supervised)	0.3981 / 0.1758 / 0.0993 / 0.0649 / 0.3674 / 0.1470 / 0.2006	0.4048 / 0.2268 / 0.2174	0.4069 / 0.5373	0.3915	0.1750	0.1299	0.0930	0.0705	0.6053
M21 - UniLM	**0.4306** / **0.2044** / **0.1218** / 0.0824 / 0.4013 / 0.1714 / 0.2228	**0.4369** / **0.2567** / **0.2483**	0.5143 / 0.6210	**0.4122**	**0.2112**	0.1455	0.0957	0.0841	0.6100
M22 - BART	**0.4416** / **0.2128** / **0.1285** / **0.0880** / **0.4100** / **0.1818** / **0.2266**	**0.4472** / **0.2646** / **0.2556**	0.5116 / 0.6215	**0.4264**	**0.2259**	0.1457	0.1037	0.0822	0.6184
M23 - Pegasus (dynamic mix)	**0.4407** / **0.2155** / **0.1307** / 0.0901 / 0.4101 / **0.1825** / **0.2260**	**0.4471** / **0.2668** / **0.2575**	0.5099 / **0.6233**	**0.4369**	**0.2283**	0.1422	0.1040	0.0797	0.6046
M23 - Pegasus (huge news)	**0.4408** / **0.2147** / **0.1295** / **0.0889** / **0.4103** / **0.1821** / **0.2273**	**0.4473** / **0.2663** / 0.2568	**0.5295** / **0.6372**	**0.4377**	**0.2286**	0.1497	0.1049	0.0845	0.6148

(a) Model scores from summarization-specific evaluation metrics.

Table (b): Model scores from other text generation evaluation metrics.

Method	BLEU	CHRF	CIDEr	METEOR	Length	Stats (cov/comp/den)	Repeated (1/2/3)
Extractive Models							
M0 - LEAD-3	11.4270	0.3892	0.2125	**0.2141**	87.4475	0.9825 / 9.6262 / 57.8001	0.2086 / 0.0310 / 0.0310
M1 - **textbf**NEUSUM	12.7784	0.3946	0.2832	**0.2183**	84.4075	0.9819 / 9.8047 / 32.8574	0.2325 / 0.0531 / 0.0531
M2 - BanditSum	12.9761	0.3897	0.3305	0.2124	78.5279	0.9836 / 10.2810 / 40.4265	0.2384 / 0.0573 / 0.0573
M3 - LATENT	12.9725	0.3897	0.3305	0.2123	78.5279	0.9834 / 10.2809 / 40.4095	0.2384 / 0.0573 / 0.0573
M4 - REFRESH	10.6568	**0.4526**	0.0677	**0.2395**	114.5684	0.9850 / 7.1059 / 53.1928	0.2127 / 0.0289 / 0.0289
M5 - RNES	11.2203	**0.4062**	0.1559	**0.2300**	99.9199	0.9938 / 7.9032 / 67.7089	0.2451 / 0.0540 / 0.0540
M6 - JECS	12.5659	**0.4310**	0.3090	0.2122	79.7797	0.9874 / 10.1111 / 26.6943	0.2041 / 0.0327 / 0.0327
M7 - STRASS	7.8330	0.3330	0.2945	0.1607	76.4859	0.9969 / 12.7835 / 59.9498	0.1864 / 0.0343 / 0.0343
Abstractive Models							
M8 - Pointer Generator	13.8247	0.3567	0.5065	0.1860	63.5211	0.9957 / 13.1940 / 26.0880	0.2015 / 0.0375 / 0.0375
M9 - Fast-abs-rl	12.9812	0.3778	0.4329	0.2014	70.8600	0.9860 / 11.0141 / **9.9859**	0.2157 / 0.0370 / 0.0370
M10 - Bottom-Up	**15.1293**	0.3523	**0.6176**	0.1887	**56.5715**	0.9811 / **14.7771** / 12.6181	**0.1856** / 0.0211 / 0.0211
M11 - Improve-abs	11.9816	0.3715	0.3356	0.2005	75.9512	**0.9674** / 10.6043 / **8.9755**	0.2499 / 0.0542 / 0.0542
M12 - Unified-ext-abs	12.8457	0.3786	0.3851	0.2017	74.4663	0.9868 / 10.7510 / 33.1106	0.2177 / 0.0493 / 0.0493
M13 - ROUGESal	13.8882	0.3668	0.4746	0.1936	66.5575	0.9853 / 13.0369 / 25.2893	0.2102 / 0.0458 / 0.0458
M14 - Multi-task (Ent + QG)	14.5276	0.3539	0.5749	0.1831	60.0294	0.9853 / **14.1828** / 22.2296	0.1985 / 0.0411 / 0.0411
M15 - Closed book decoder	13.4158	0.3675	0.4648	0.1925	68.2858	0.9866 / 12.0588 / 27.3686	0.2074 / 0.0444 / 0.0444
M16 - SENECA	13.7676	0.3660	0.5233	0.1966	64.9710	0.9866 / 12.3610 / 16.7640	0.2146 / 0.0303 / 0.0303
M17 - T5	**19.3891**	0.3833	**0.7763**	0.2140	59.5288	**0.9775** / **14.2002** / 12.9565	**0.1810** / 0.0209 / 0.0209
M18 - NeuralTD	12.9241	0.3783	0.3543	0.2038	74.4033	0.9830 / 10.7768 / **12.4443**	0.2645 / 0.0901 / 0.0901
M19 - BertSum-abs	14.9525	0.3649	**0.6240**	0.1876	60.8893	0.9517 / **13.9197** / **12.3254**	**0.1697** / 0.0156 / 0.0156
M20 - GPT-2 (supervised)	13.9364	0.3678	0.5787	0.1759	51.8352	0.9791 / **15.9839** / 15.4999	0.1875 / 0.0362 / 0.0362
M21 - UniLM	**15.5736**	**0.4230**	0.5294	0.2084	67.1960	**0.9685** / 11.5672 / **11.7908**	**0.1722** / 0.0180 / 0.0180
M22 - BART	**17.1005**	**0.4271**	0.7573	0.2105	62.2989	**0.9771** / 12.8811 / 15.2999	**0.1627** / 0.0127 / 0.0127
M23 - Pegasus (dynamic mix)	**18.6517**	**0.4261**	0.7280	0.2131	64.1348	0.9438 / 13.7208 / 11.6003	0.1855 / 0.0355 / 0.0081
M23 - Pegasus (huge news)	**17.8102**	0.3912	**0.6595**	**0.2189**	66.7559	0.9814 / 12.9473 / 14.9850	0.1883 / 0.0251 / 0.0251

(b) Model scores from other text generation evaluation metrics.

Table 4: Model scores from automatic evaluation metrics available in the evaluation toolkit. The five highest scores for each metric (and lowest for Length and Repeated-1/2/3) are bolded.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. 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 3110–3120, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1307

Léo Bouscarrat, Antoine Bonnefoy, Thomas Peel, and Cécile Pereira. 2019. STRASS: A light and effective method for extractive summarization based on sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 243–252, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-2034

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. 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. **DOI:** `https://doi.org/10.18653/v1/P18-1060`

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/P18-1063`

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. **DOI:** `https://doi.org/10.18653/v1/P19-1264`

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, Portorož, Slovenia. European Language Resources Association (ELRA).

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 update summarization task. In *TAC*.

Hoa Trang Dang and Karolina Owczarzak. 2009. Overview of the TAC 2009 summarization track. In *Proceedings of the Text Analysis Conference*.

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).

Daniel Deutsch and Dan Roth. 2020. SacreROUGE: An Open-Source Library for Using and Developing Summarization Evaluation Metrics. **DOI:** `https://doi.org/10.18653/v1/2020.nlposs-1.17`

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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13042–13054.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1409`, **PMID:** 30577265

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/2020.acl-main.454`

Kavita Ganesan. 2015. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1347–1354. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/2020.acl-main.124`

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. **DOI:** `https://doi.org/10.18653/v1/D18-1443`

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D15-1013`, **PMID:** 24802104

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. **DOI:** `https://doi.org/10.18653/v1/N18-1065`

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 687–697, Melbourne, Australia. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/P18-1064`, **PMCID:** PMC6428206

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/P19-1330`

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

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

Yichen Jiang and Mohit Bansal. 2018. Closed-book training to improve summarization encoder memory. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4067–4077, Brussels, Belgium. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1440`

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. **DOI:** `https://doi.org/10.18653/v1/D18-1208`

404

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Retrieved from `https://repository.upenn.edu/asc_papers/43`.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/2020.emnlp-main.750`

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1207`

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics. **DOI:** `https://doi.org/10.3115/1626355.1626389`

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer.

2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*. **DOI:** `https://doi.org/10.18653/v1/2020.acl-main.703`

Chin-Yew Lin. 2004a. Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough? In *NTCIR*.

Chin-Yew Lin. 2004b. 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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D19-1387`

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300. **DOI:** `https://doi.org/10.1162/COLI_a_00123`

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/2020.acl-main.173`

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and

their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv: 1602.06023*. **DOI:** `https://doi.org/10.18653/v1/K16-1028`

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. 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 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/N18-1158`

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Karolina Owczarzak, Peter A. Rankel, Hoa Trang Dang, and John M. Conroy. 2012. Assessing the effect of inconsistent assessors on summarization evaluation. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8–14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 359–362. The Association for Computer Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. **DOI:** `https://doi.org/10.3115/1073083.1073135`

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653, New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/N18-2102`

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Maxime Peyrard. 2019. 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. **DOI:** `https://doi.org/10.18653/v1/P19-1502`

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/W17-4510`

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/W15-3049`

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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 e-prints*.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D19-1320`

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.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. A graph-theoretic summary evaluation for ROUGE. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 762–767, Brussels, Belgium. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1085`

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive 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 3280–3291, Hong Kong, China. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D19-1323`, **PMID:** 31698456, **PMCID:** PMC6855099

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information processing Systems*, pages 3104–3112.

Oleg V. Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: human-free quality estimation of document summaries. *CoRR*, abs/2002.09836. **DOI:** `https://doi.org/10.18653/v1/2020.eval4nlp-1.2`

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575. **DOI:** `https://doi.org/10.1109/CVPR.2015.7299087`

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/2020.acl-main.450`, **PMCID:** PMC7367613

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256. **DOI:** `https://doi.org/10.1007/BF00992696`

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In *EMNLP-IJCNLP 2019*, pages 3292–3303, Hong Kong, China. Association for Computational Linguistics.

Fangfang Zhang, Jin-ge Yao, and Rui Yan. 2018a. 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. **DOI:** `https://doi.org/10.18653/v1/D18-1089`

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

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*.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018b. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1088`

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019b. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Lin-*

guistics, pages 5059–5069, Florence, Italy. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/P19-1499`, **PMID:** 31638247, **PMCID:** PMC6854546

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *EMNLP-IJCNLP 2019*, pages 563–578, Hong Kong, China. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D19-1053`

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 447–454, New York City, USA. Association for Computational Linguistics. **DOI:** `https://doi.org/10.3115/1220835.1220892`

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *ACL 2018*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/P18-1061`

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

9 Appendix

Data Collection   The data collection interface used by both crowd-source and expert annotators is presented in Figure 3. In the annotation process, judges were first asked to carefully read the content of the source article and next proceed to evaluating the associated summaries along four axes: *relevance*, *consistency*, *fluency*, and *coherence*.

Instructions

In this task you will evaluate the quality of summaries written for a news article.
To correctly solve this task, follow these steps:

1. Carefully read the news article, be aware of the information it contains.
2. Read the proposed summaries A-F (6 in total).
3. Rate each summary on a scale from **1** (worst) to **5** (best) by its *relevance*, *consistency*, *fluency*, and *coherence*.

Definitions

Relevance:

The rating measures how well the summary captures the key points of the article.
Consider whether all and only the important aspects are contained in the summary.

Consistency:

The rating measures whether the facts in the summary are consistent with the facts in the original article.
Consider whether the summary does reproduce all facts accurately and does not make up untrue information.

Fluency

This rating measures the quality of individual sentences, are they well-written and grammatically correct.
Consider the quality of individual sentences.

Coherence:

The rating measures the quality of all sentences collectively, to the fit together and sound naturally.
Consider the quality of the summary as a whole.

Article

${article}

Summaries

Summary A

${grounding}

Relevance	1	2	3	4	5
Consistency	1	2	3	4	5
Fluency	1	2	3	4	5
Coherence	1	2	3	4	5

Figure 3: Example of the data collection interface used by crowd-source and expert annotators.