SummEval: Re-evaluating Summarization Evaluation

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, Dragomir Radev

CNeRG Reading Group Presentation
(9th September, 2021)

Kiran Purohit
Rajdeep Mukherjee

Summarization Task

- Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks).
  - Mani and Maybury, 1999

- CNN/DailyMail (CNNDM) corpus (Hermann et al., 2015, Nallapati et al., 2016)
  - News articles and bullet point summaries.
  - Standard dataset for training summarization models

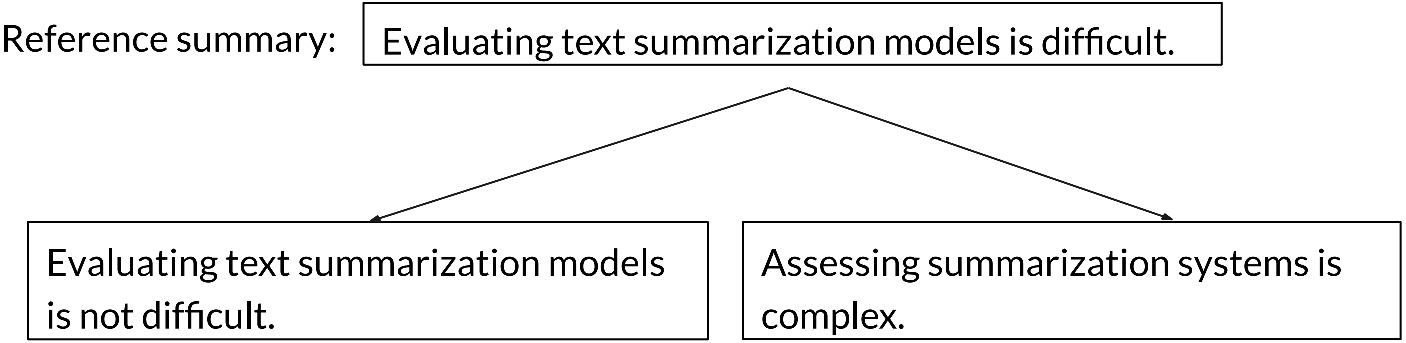
Summarization Evaluation

- ROUGE (Lin, 2004): lexical overlap with a reference summary (or summaries)
- E.g., ROUGE-N Recall:

$$\frac{\sum_{S\in\{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S\in\{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}$$

Summarization Evaluation

- ROUGE (Lin, 2004): lexical overlap with a reference summary (or summaries)



Reference summary: Evaluating text summarization models is difficult.

Evaluating text summarization models is not difficult.

Assessing summarization systems is complex.

ROUGE-1: 92.31

ROUGE-1: 36.36

Observations from Prior Research - Metric Evaluation

- Effects of inconsistencies in human annotations on the rankings of evaluated summarization systems (Owczarzak et al. 2012) - 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.
- Analyzing different variants of the ROUGE metric (Rankel et al. 2013, Graham 2015) - Higher-order and less commonly reported ROUGE settings showed a higher correlation with human judgments.
- 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.

Observations from Prior Research - Dataset Evaluation

- (Dernoncourt et al. 2018) - differences in formats of available corpora.
- (Krysci´nski 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.

Observations from Prior Research - Model Evaluation

- (Zhang et al. 2018) - word-level extractive models achieved a similar level of abstraction to fully abstractive models.
- (Kedzie et al. 2018) - In the current setting, the training signal is dominated by biases present in summarization datasets preventing models from learning accurate content selection.
- (Krysci´nski et al. 2020, Maynez et al. 2020) showed that
  - The issue of hallucinating facts touches up to 30% of generated summaries - Poor factual correctness.
  - Improving factual faithfulness is a critical issue in summarization.
  - Current available evaluation methods, such as ROUGE and BertScore, are not sufficient to study the problem at hand.

Need for Consistent Summarization Evaluation

- Metric evaluation
  - New metrics proposed but not widely adopted
  - Metrics are evaluated on DUC and TAC shared tasks, not representative of modern systems (Peyrard, 2019)

- Model evaluation
  - Recent papers vastly differ in their evaluation protocol as noted in Hardy et al. (2019)
  - Most papers compare to only several other models

Major Contributions

- The authors 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.
- The authors consistently benchmark 23 recent summarization models using the aforementioned automatic evaluation metrics.
- The authors 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.
- The authors release a toolkit of 14 evaluation metrics with an extensible and unified API to promote the reporting of additional metrics in papers.
- The authors 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.

Re-evaluating Metrics and Models

- Re-evaluating metrics
    - 14 automatic evaluation metrics
    - Toolkit with extensible and unified API
    - Largest and most diverse, in terms of model types, collection of human judgments of model-generated summaries on the CNNDM dataset

- Re-evaluating models
    - Consistently benchmark 23 recent summarization models
    - Largest collection of summaries on the CNNDM news dataset for easier comparison

Evaluation Metrics

- **ROUGE-based:** ROUGE (Lin, 2004b); ROUGE-WE (Ng and Abrecht, 2015); S3 (Peyrard et al., 2017)
- **Contextual Embedding-Based:** BertScore (Zhang* et al., 2020), MoverScore (Zhao et al., 2019), Sentence Mover's Similarity (Clark et al., 2019); SummaQA (Scialom et al., 2019)
    - **Reference-less:** BLANC (Vasilyev et al., 2020); SUPERT (Gao et al., 2020)
- **Machine translation, text generation metrics:** BLEU (Papineni et al., 2002); CHRF (Popović, 2015); METEOR (Lavie and Agarwal, 2007); CIDEr (Vedantam et al., 2015)
- **Data Statistics:** Grusky et al. (2018)

Evaluation Metrics - ROUGE-Based:

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.

Contextual Embedding-Based:

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) A BERT-based question-answering model is applied 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.

Reference-less:

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.

Machine translation, Text Generation metrics:

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.

Summarization Models

- 23 models introduced from 2017 to 2019

Summarization Models

Extractive Models	
NEUSUM (Zhou et al., 2018)	BanditSum (Dong et al., 2018)
LATENT (Zhang et al., 2018b)	REFRESH (Narayan et al., 2018)
RNES (Wu and Hu, 2018)	JECS (Xu and Durrett, 2019)
STRASS (Bouscarrat et al., 2019)	

Summarization Models

Non-pretrained Abstractive Models	
Pointer Generator (See et al., 2017)	ROUGESal (Pasunuru and Bansal, 2018)
Fast-abs-rl (Chen and Bansal, 2018)	Multi-task (Ent + QG) (Guo et al., 2018)
Bottom-Up (Gehrmann et al., 2018)	Closed book decoder (Jiang and Bansal, 2018)
Improve-abs (Kryściński et al., 2018)	SENECA (Sharma et al., 2019)
Unified-ext-abs (Hsu et al., 2018)	NeuralTD (Böhm et al., 2019)

Summarization Models

Pretrained Abstractive Models	
T5 (Raffel et al., 2019)	UniLM (Dong et al., 2019)
BertSum-abs (Liu and Lapata, 2019)	BART (Lewis et al., 2019)
GPT-2 (Ziegler et al., 2019)	Pegasus (Zhang et al., 2019a)

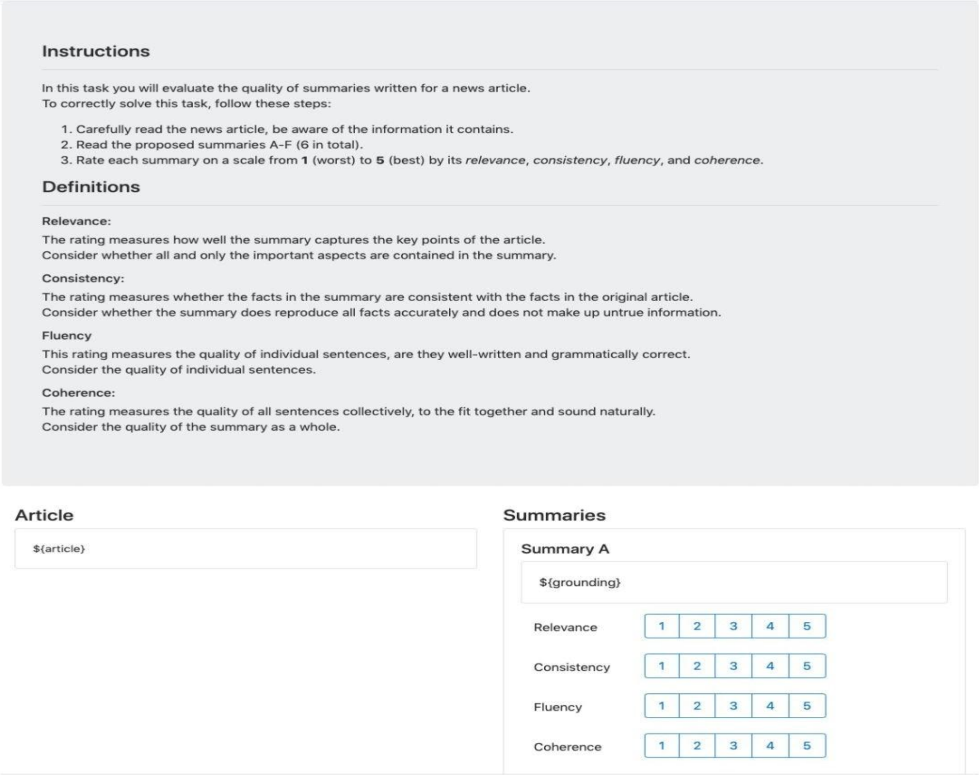
Human Judgments

- 100 articles from the CNN/DM dataset; 16 models; 3 expert and 5 crowdsourced judgments

- 4 quality dimensions (rated from 1 to 5, higher better)
  - **Coherence -** the structure and organization of all summary sentences
  - **Consistency -** the factual alignment between summary and input
  - **Fluency -** the grammatical quality of individual sentences
  - **Relevance -** selection of important content from the source.

- Two rounds of expert annotations for better agreement (0.71 Krippendorf's alpha)

Annotation Interface



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

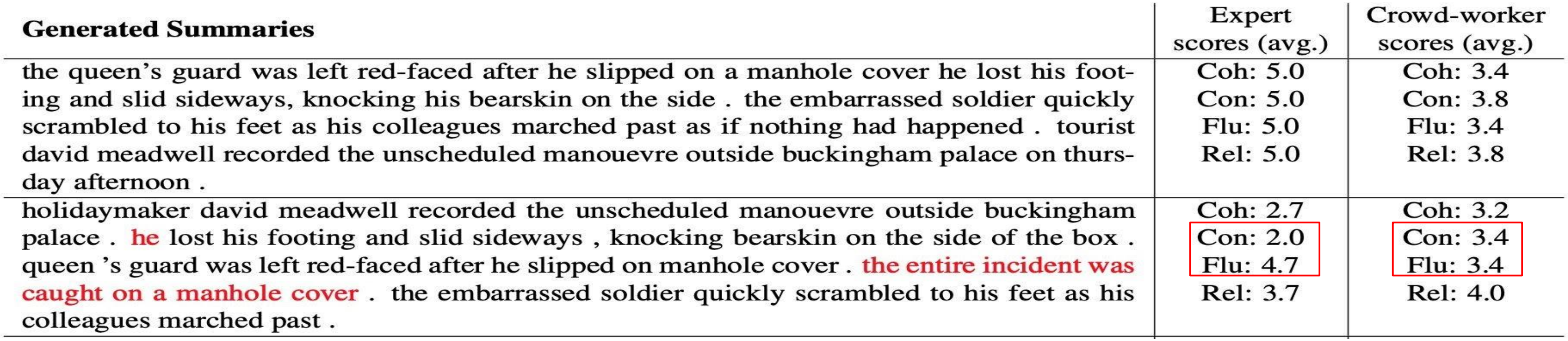
${article}

Summaries

Summary A

${grounding}

Problems with Crowdsourced Judgments

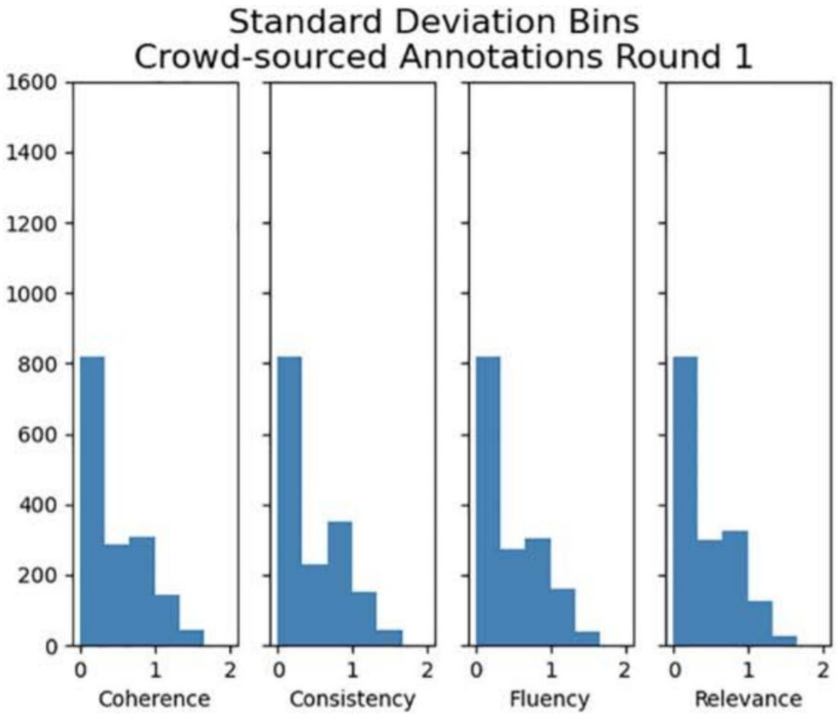


Generated Summaries	Expert scores (avg.)	Crowd-worker scores (avg.)
the queen's guard was left red-faced after he slipped on a manhole cover he lost his footing and slid sideways, knocking his bearskin on the side . the embarrassed soldier quickly scrambled to his feet as his colleagues marched past 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

- Failure of crowdsourced workers to differentiate among dimensions.

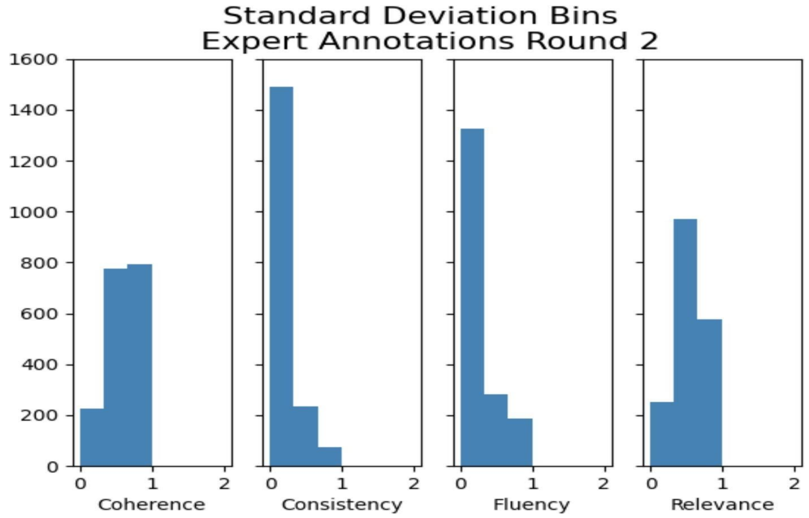
Problems with Crowdsourced Judgments

- Insufficient distinction made by the annotators between the 4 scored dimensions.
- Overall quality of a summary biased scores of the individual dimensions.



Agreement across Dimensions among Experts

- Strong agreement for consistency and fluency.
- Coherence and relevance agreement point to subjectivity of dimensions and need for finer-grained instructions.

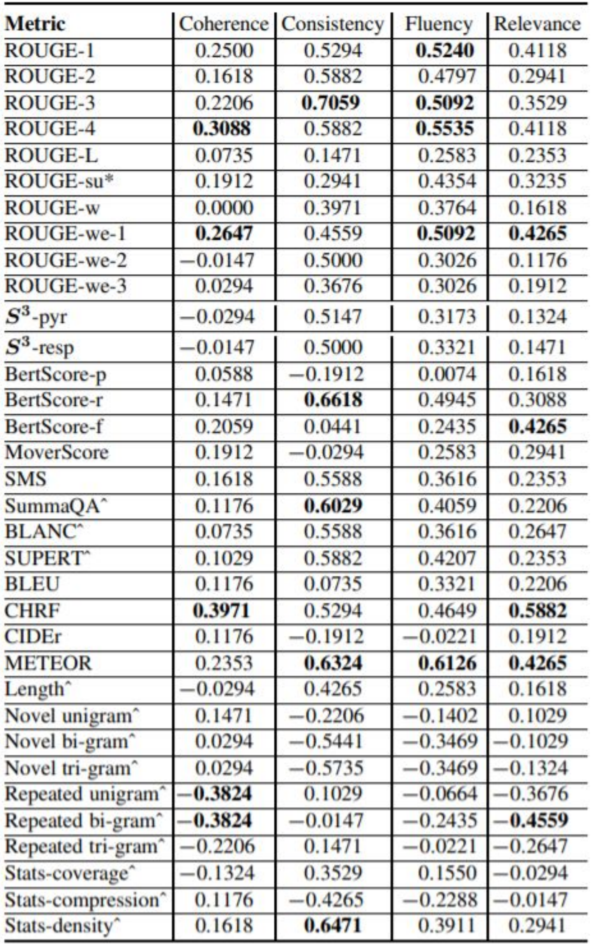


Metric Re-evaluation

- Kendall’s tau rank for system-level ranking as in Louis and Nenkova, 2013

Metric Re-evaluation

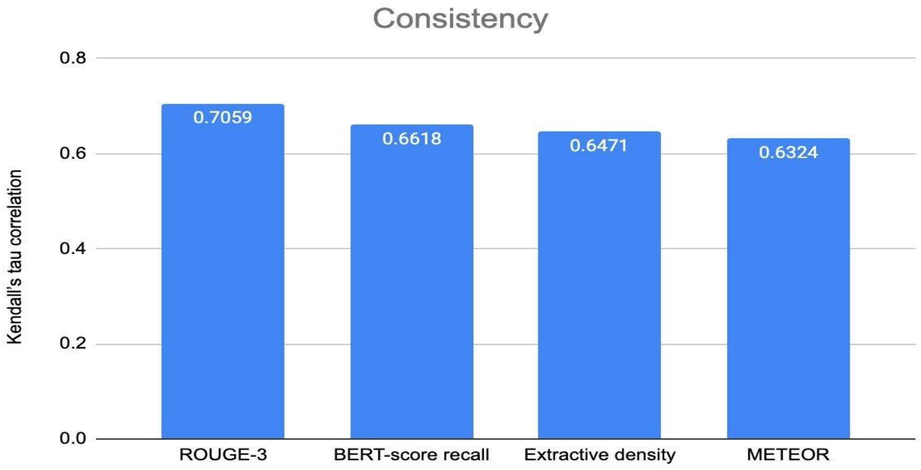
- Most metrics have the lowest correlation within the coherence dimension.
- Low and moderate correlation scores along the relevance dimension - inherent subjectiveness of the dimension and the difficulty of collecting consistent human annotations.
- Strong correlation with consistency:
  - Low abstractiveness of most neural models.
  - High inter-annotator agreement between expert judges.
- Higher correlation between all evaluated dimensions and ROUGE scores computed for higher-order n-grams in comparison to ROUGE-L.



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
SummaQA^	0.1176	**0.6029**	0.4059	0.2206
BLANC^	0.0735	0.5588	0.3616	0.2647
SUPERT^	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**
Length^	−0.0294	0.4265	0.2583	0.1618
Novel unigram^	0.1471	−0.2206	−0.1402	0.1029
Novel bi-gram^	0.0294	−0.5441	−0.3469	−0.1029
Novel tri-gram^	0.0294	−0.5735	−0.3469	−0.1324
Repeated unigram^	**−0.3824**	0.1029	−0.0664	−0.3676
Repeated bi-gram^	**−0.3824**	−0.0147	−0.2435	**−0.4559**
Repeated tri-gram^	−0.2206	0.1471	−0.0221	−0.2647
Stats-coverage^	−0.1324	0.3529	0.1550	−0.0294
Stats-compression^	0.1176	−0.4265	−0.2288	−0.0147
Stats-density^	0.1618	**0.6471**	0.3911	0.2941

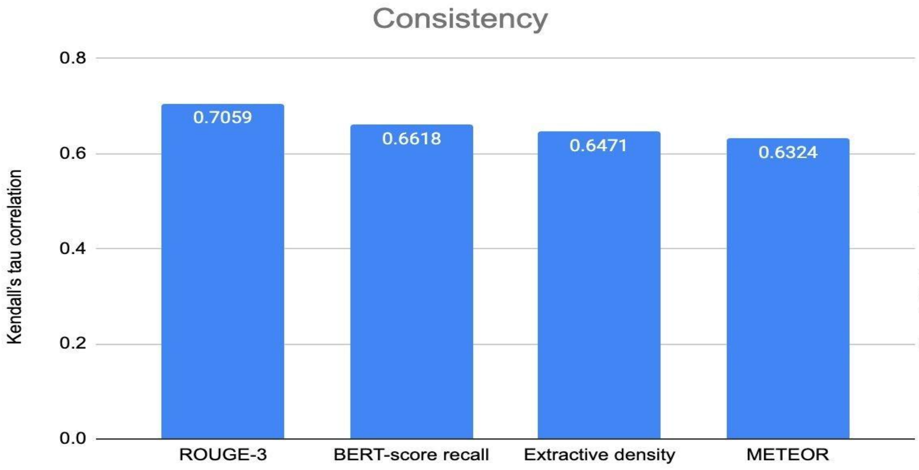
Metric Re-evaluation

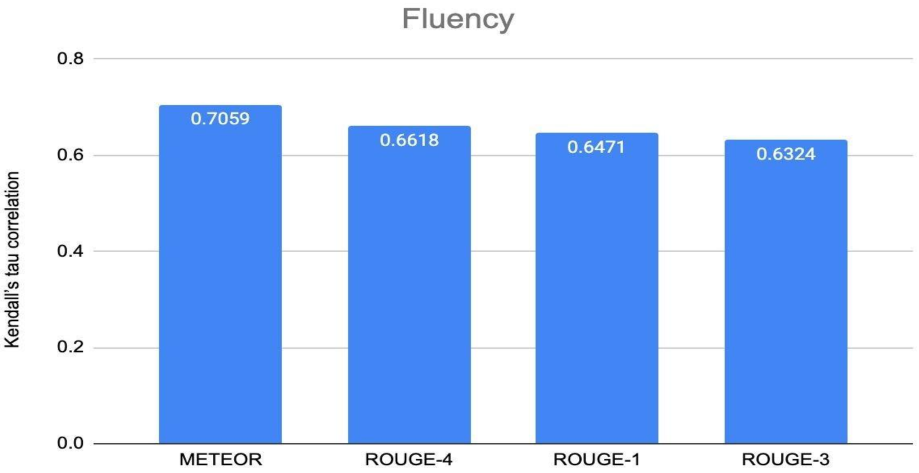
- Strong correlation with consistency, perhaps due to **extractive** nature of dataset



Metric Re-evaluation

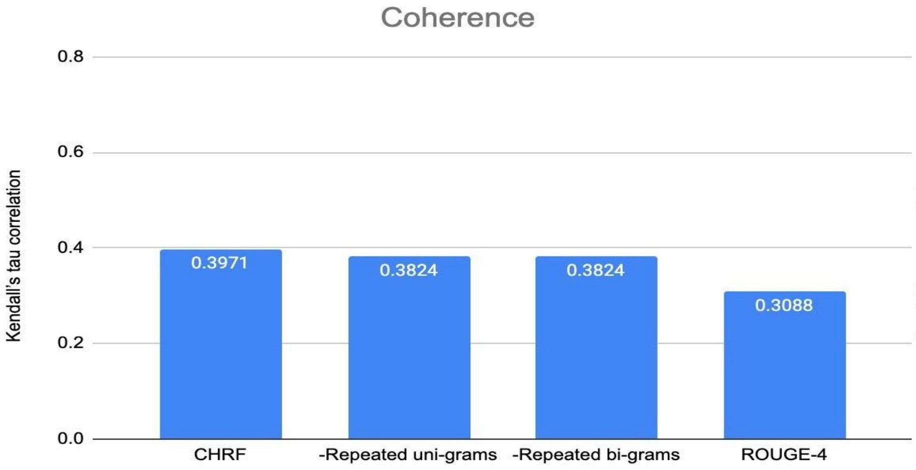
- Stronger correlations in general over these dimensions.

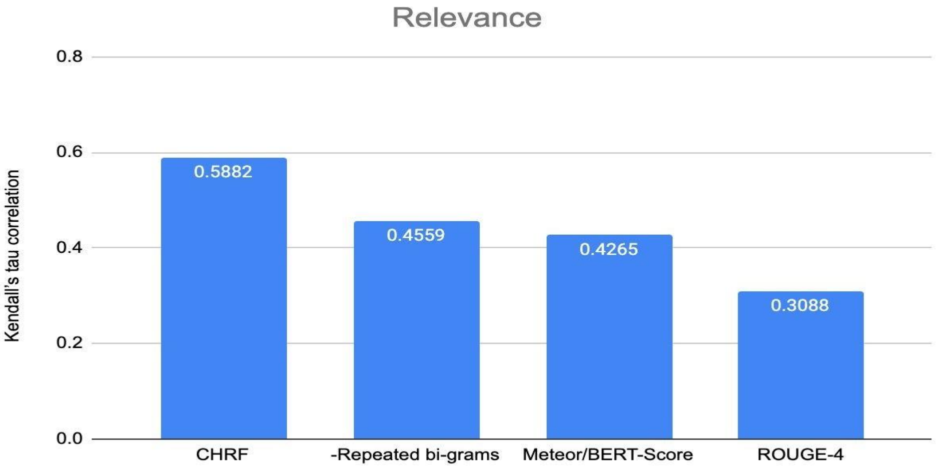


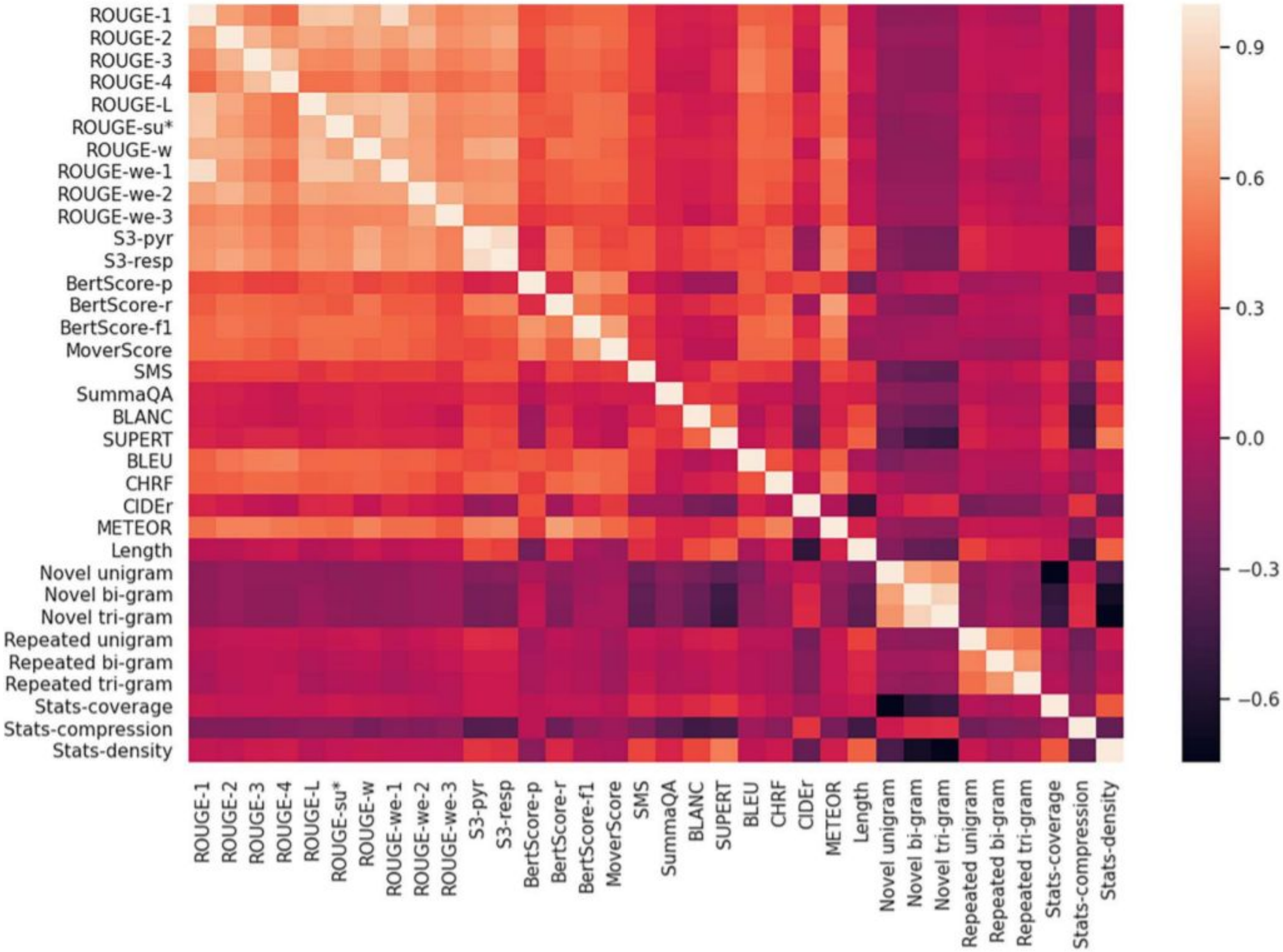


Metric Re-evaluation

- Weaker correlations potentially to inherent subjectiveness of the dimension and the difficulty of collecting consistent human annotations.

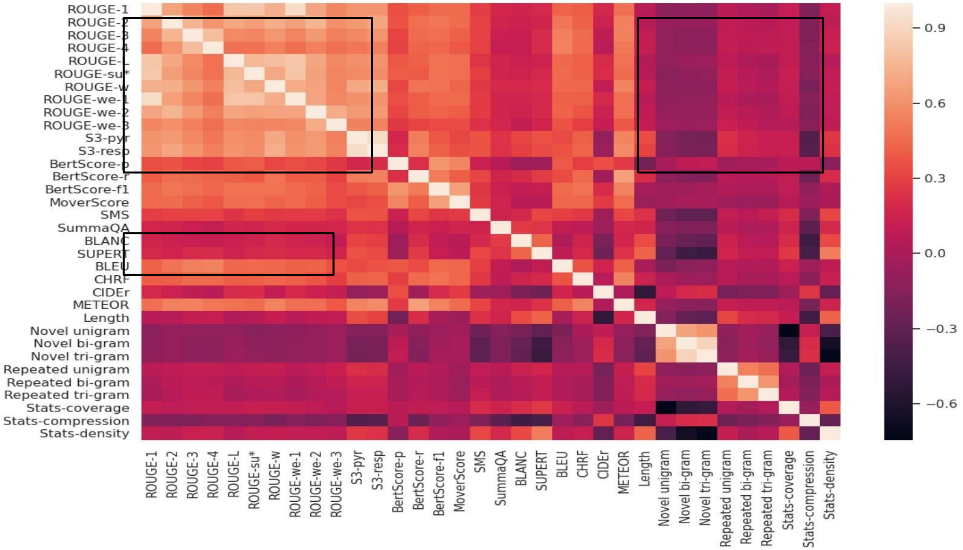


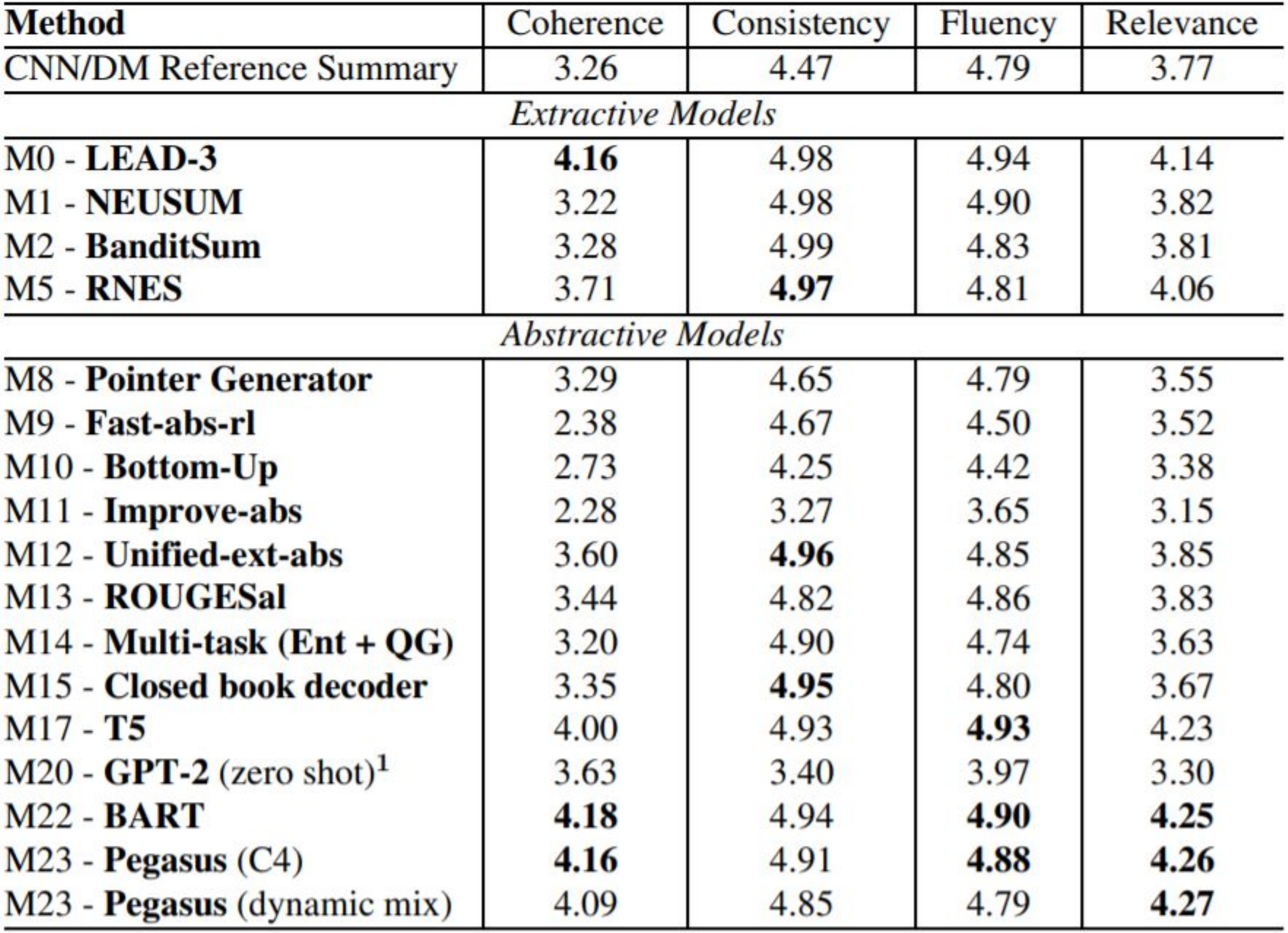



ROUGE-1
ROUGE-2
ROUGE-3
ROUGE-4
ROUGE-L
ROUGE-su*
ROUGE-w
ROUGE-we-1
ROUGE-we-2
ROUGE-we-3
S3-pyr
S3-resp
BertScore-p
BertScore-r
BertScore-f1
MoverScore
SMS
SummaQA
BLANC
SUPERT
BLEU
CHRF
CIDEr
METEOR
Length
Novel unigram
Novel bi-gram
Novel tri-gram
Repeated unigram
Repeated bi-gram
Repeated tri-gram
Stats-coverage
Stats-compression
Stats-density
0.9
0.6
0.3
0.0
−0.3
−0.6

Metric Re-evaluation

- Strong correlations among lexical overlap metrics.
- Novelty and repetitiveness show a weak negative correlation with ROUGE-related metrics.
- Weak correlation of reference-less metrics with most other evaluated metrics.

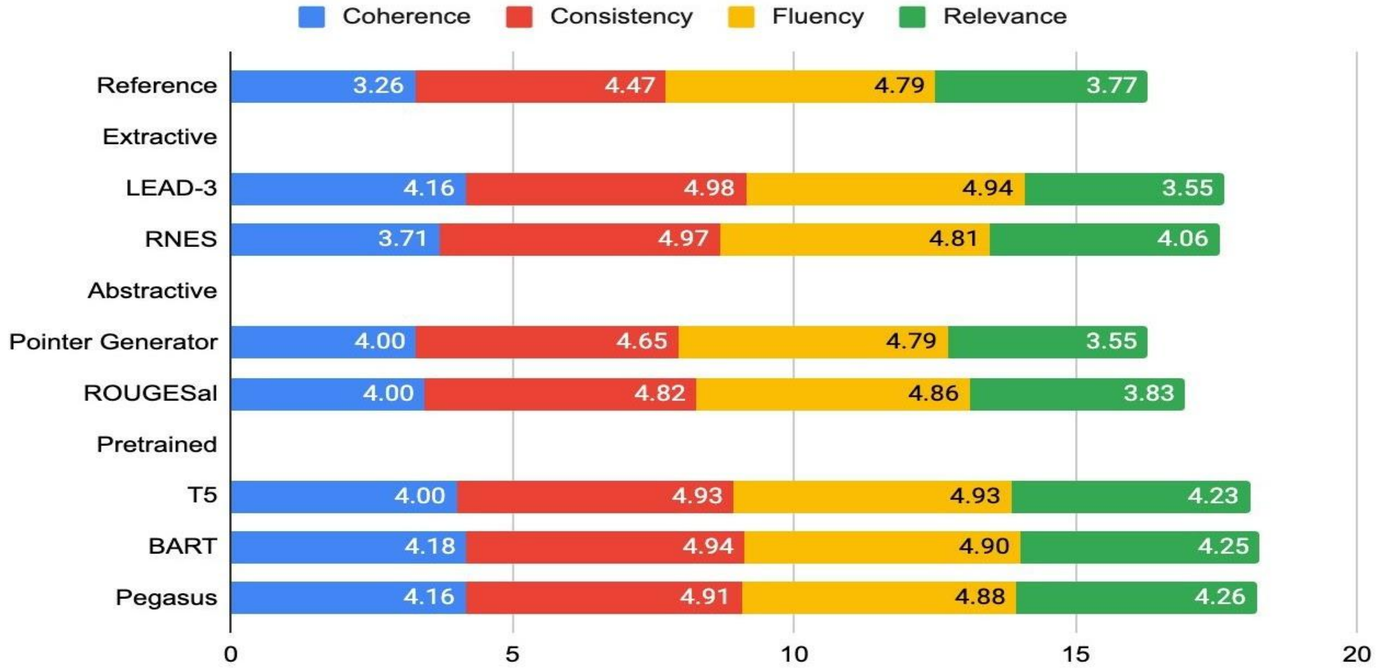




Method	Coherence	Consistency	Fluency	Relevance
CNN/DM Reference Summary	3.26	4.47	4.79	3.77
Extractive Models				
M0 - **LEAD-3**	**4.16**	4.98	4.94	4.14
M1 - **NEUSUM**	3.22	4.98	4.90	3.82
M2 - **BanditSum**	3.28	4.99	4.83	3.81
M5 - **RNES**	3.71	**4.97**	4.81	4.06
Abstractive Models				
M8 - **Pointer 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**

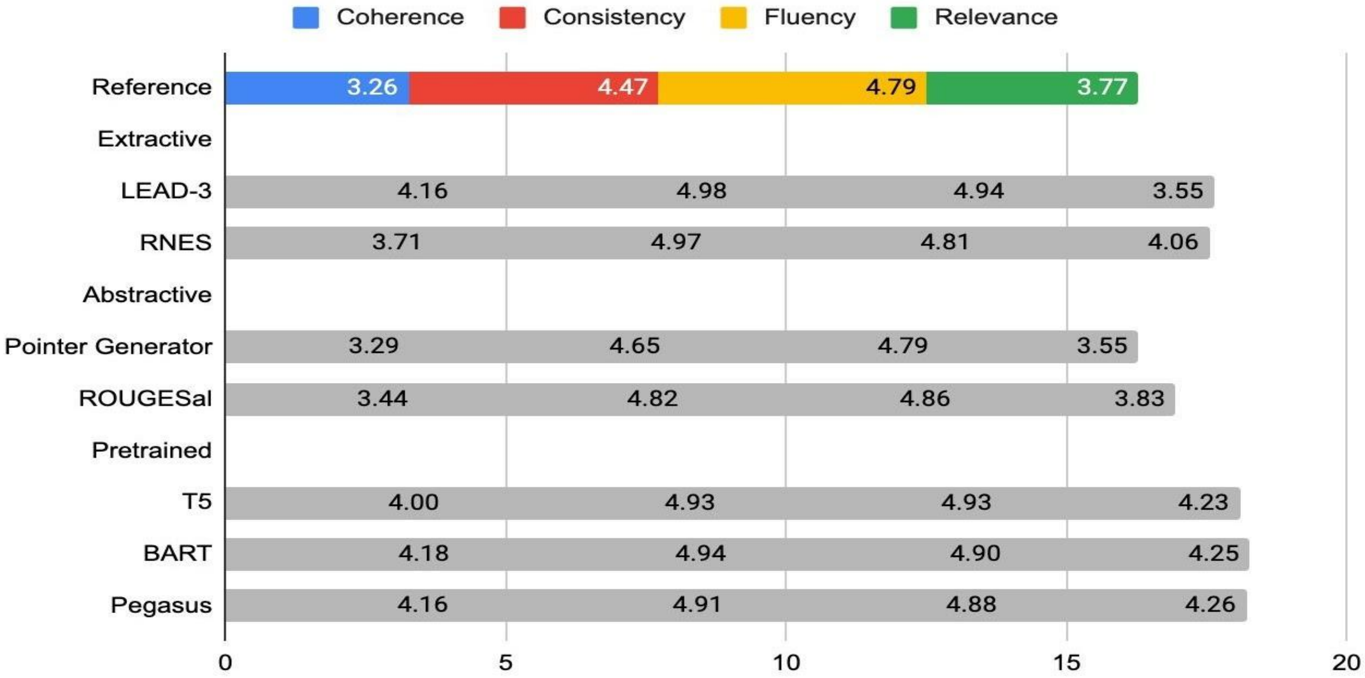
Model Re-evaluation

- Scores from 1 to 5 (best).

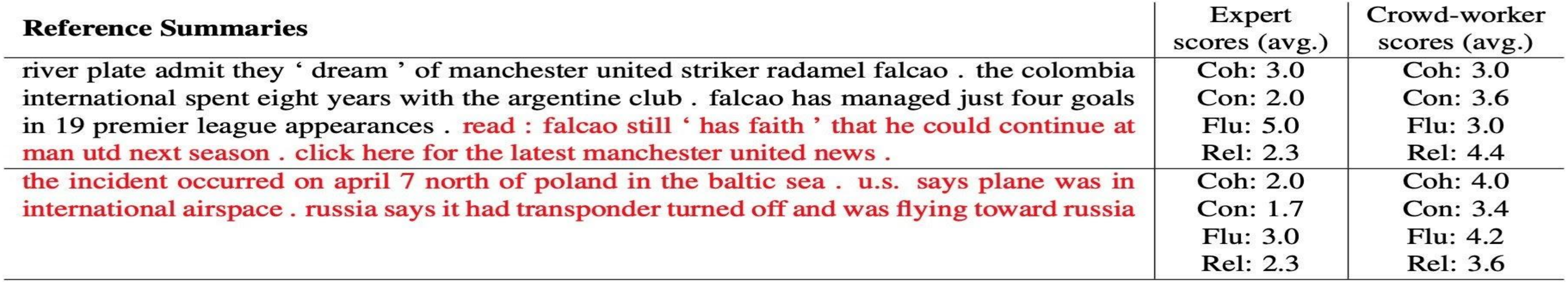


Model Re-evaluation

- Reference summaries are far from ideal.



Model Re-evaluation: Reference Summary Scores

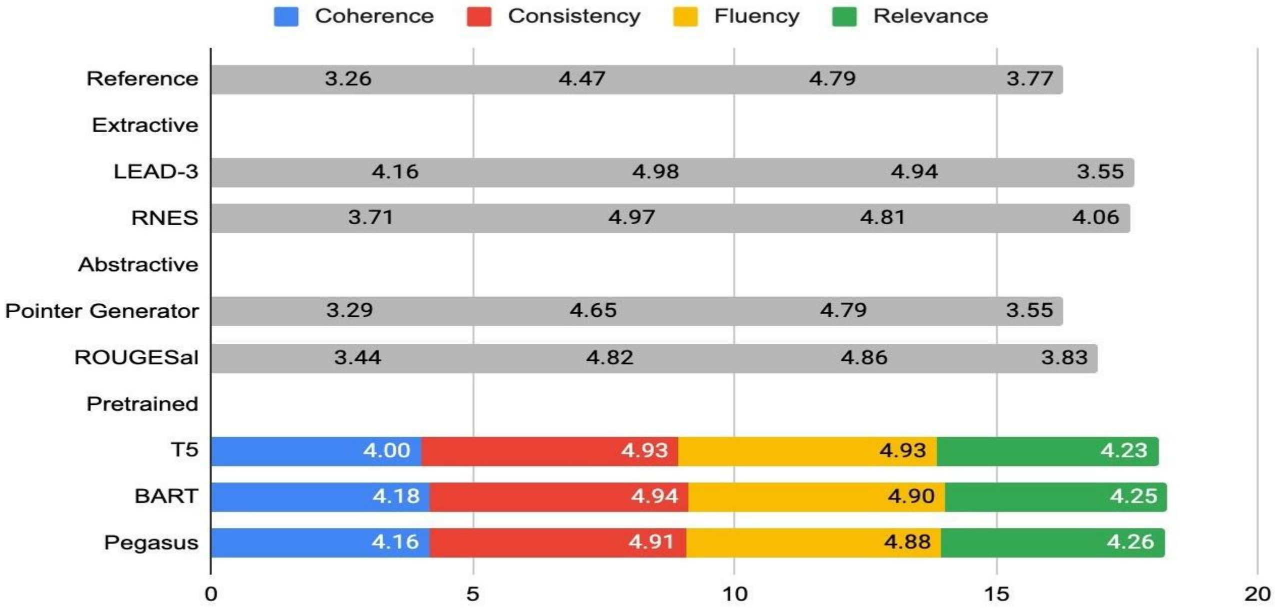


Reference Summaries	Expert scores (avg.)	Crowd-worker scores (avg.)
river plate admit they ‘ dream ’ of manchester united striker radamel falcao . the colombia international spent eight years with the argentine club . falcao has managed just four goals in 19 premier league appearances . read : falcao 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

- Errors in consistency, relevance and fluency result from automatic dataset construction

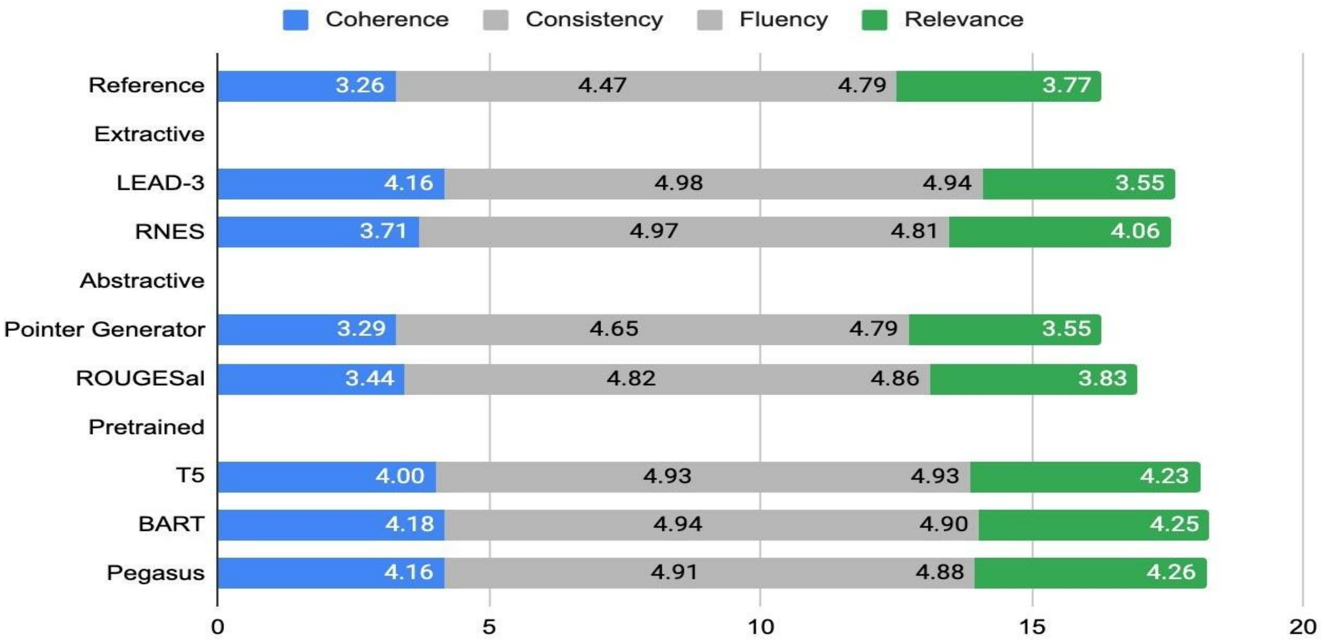
Model Re-evaluation

- Reference summaries are far from ideal.
- Improvements with pretrained models.



Model Re-evaluation

- Coherence and relevance can still be improved on this dataset.



SummEval Toolkit

- Install

```
% pip install summ-eval
```

- Import

```python
from summ_eval.rouge_metric import RougeMetric
rouge = RougeMetric()
```

- Evaluate!

```python
summaries = ["This is one summary","This is another summary"]
references = ["This is one reference", "This is another"]
rouge_dict = rouge.evaluate_batch(summaries, references)
```

Summary

- Re-evaluated 14 automatic metrics and 23 summarization models.
- Released an evaluation toolkit, https://github.com/Yale-LILY/SummEval, along with expert and crowdsourced human judgments across 4 quality dimensions.
- We promote a more comprehensive comparison of summarization models

Thank you!