A Study on Summarizing and Evaluating Long Documents

Anonymous ACL submission

Abstract

Text summarization has been a key language generation task for over 60 years. The field has advanced considerably during the past two years, benefiting from the proliferation of pre-trained Language Models (LMs). However, the field is constrained by two factors: 1) the absence of an effective automatic evaluation metric and 2) a lack of effective architectures for long document summarization. Our first contribution is to demonstrate that a set of semantic evaluation metrics (BERTScore, MoverScore and our novel metric, BARTScore) consistently and significantly outperform ROUGE. Using these metrics, we then show that combining transformers with sparse self-attention is a successful method for long document summarization and is competitive with the state of the art. Finally, we show that sparsifying self-attention does not degrade model performance when using transformers for summarization.

1 Introduction

Summaries play a key role in communicating written information. Defined as a document reduced only to its essential content, a summary helps readers understand material more easily and quickly. In a digital world with ever-increasing volumes of written content online, summaries play a central role in synthesising information into a digestible format for readers. For example, between the onset of Covid-19 in January and May 2020 there were an estimated 23,000 research papers published on the virus with the number doubling every week[1].

Text summarization has been researched for over 60 years (Saggion and Poibeau, 2013) but has seen dramatic progress over the past five years since the introduction of the *seq2seq* neural paradigm (Sutskever et al., 2014). This used a recurrent neural network (RNN) (Rumelhart et al., 1986) to encode the input sequence into a single vectorial representation and another RNN to extract the target sequence. This approach allowed the generation of sequences of arbitrary length conditioned upon an input document and therefore was adopted by the summarization community as the first viable attempt at abstractive summarization (e.g. (See et al., 2017), (Nallapati et al., 2016)). The pace of progress further accelerated leveraging transformers (Vaswani et al., 2017) as these are able to generate outputs with higher fluency and coherence than was previously possible. This improvement in performance drove a push into a wider and more diversified range of datasets, broadening from summarizing short news articles (e.g. (Hermann et al., 2015), (Narayan et al., 2018)) to generating news headlines (Rush et al., 2015), longer scientific documents (Cohan et al., 2018) and multiple documents (Fabbri et al., 2019).

A challenge in the field is how to approach long document summarization (generally defined as over 1K tokens) as the dominant transformer architectures become inefficient when using long sequences. This is one focus of this study. The other main focus is on the evaluation metrics as these have been neglected and are an essential yardstick for measuring progress. There are a number of weaknesses with the prevailing ROUGE (Lin, 2004) metrics, creating a reliance on human evaluation which is expensive, impractical and opaque. This study first establishes a superior set of evaluation metrics and uses these to provide an analysis of long document summarization architectures. Our **main contributions** are:

- We develop a novel text generation evaluation metric, BARTScore. This correlates approximately 2x more strongly with human judgement than ROUGE and performs competitively or better than all metrics we tested.
- Through novel and rigorous experimentation, we establish a superior set of model-

[1] `https://tinyurl.com/ybmmdjkl`.

based evaluation metrics for summarization – BARTScore, BERTScore, Mover-1 and Mover-2. All significantly outperform ROUGE on five tasks spanning three datasets.

- We demonstrate that sparsifying self-attention does not significantly harm summarization models' performances.
- We achieve summarization performance on par with state of the art on the arXiv dataset using relatively modest computational resources.

2 Related Work

2.1 Summarization Models

The most successful approaches to summarization use the transformer encoder-decoder (TED) architecture Vaswani et al. (2017). A selection are outlined below.

BART A denoising autoencoder trained to recover the original text after it has been corrupted by a noising function. Lewis et al. (2019) investigate a number of corrupting techniques and find two best-performing approaches: 1) recovering the order of shuffled sentences in a document; 2) in-filling spans of masked tokens.

PEGASUS A TED with a novel pre-training procedure, Gap-Sentences Generation (Zhang et al., 2019a). This identifies important sentences within the document and predicts these conditioned on the remainder of the document.

ProphetNet Yan et al. (2020) recognise that language generation models overfit on local correlations at the expense of global correlations due to training using teacher-forcing (Williams and Zipser, 1989). They modify the task to predict n tokens ahead, hence altering the *seq2seq* objective from predicting $p(y_t|y^i_{<t}, x)$ into predicting $p(y_{t:t+n-1}|y^i_{<t}, x)$.

Longformer The above models are well-suited to short documents but ill-suited to long documents due to the quadratic memory complexity of self-attention with respect to the input length. Hence, their memory requirements become too large for standard GPUs on longer sequences. Beltagy et al. (2020) propose sparsifying self-attention to address this shortcoming. Their sliding window attention reduces memory complexity to linear as each token attends only to a fixed number each side.

LED The Longformer was originally designed as a transformer encoder, vis-a-vis BERT (Devlin et al., 2018). Beltagy et al. (2020) later combine sliding window attention with BART to create the Longformer Encoder Decoder (LED), a TED suited to long document summarization. The LED is created using three steps: 1) copy the self-attention weights from BART's encoder into the Longformer's corresponding sliding window attention layers; 2) replace the self-attention layers in BART's encoder with the Longformer's corresponding self-attention layers; 3) widen BART's positional embedding matrix to the desired maximum input length. The authors do not experiment with this model so we investigate this in this study.

Other Approaches Reformer (Kitaev et al., 2020) uses the LSH algorithm to reduce the memory complexity of self-attention to log-linear. The Linformer (Wang et al., 2020) approximates the self-attention computation using a low-rank matrix, also reducing the complexity to linear. At the time of writing, BIGBIRD (Zaheer et al., 2020) is SOTA for long document summarization using sparse self-attention, which combines random attention patterns with global attention for selected tokens. Approximate self-attention has received considerable interest recently, although there has not been any systematic comparison between the variants at the time of writing.

2.2 Evaluation Metrics

ROUGE The prevailing automatic evaluation metric for text summarization (Lin, 2004). Computes a score as a function of the co-occurrences of n-grams between a candidate and reference summary. We report ROUGE-1,2 and L following convention. The former compute the co-occurrences of 1 / 2-grams, while ROUGE-L measures the longest common subsequence between two texts. ROUGE has several drawbacks. For example, it performs a surface-level comparison which penalises lexically diverse but semantically equivalent texts. Consequently ROUGE has been shown to correlate poorly with human judgement (Böhm et al., 2019).

We contrast ROUGE with a set of more recent, model-based metrics, outlined below. These evaluate the similarity between a target and hypothesis sequence in an embedded space and use transformers to compute the contextual representations. These were developed for machine translation evaluation to better reflect semantics than surface-level metrics (e.g. BLEU (Papineni et al., 2002)). We postulate that this property will also make them useful for automatic summarization evaluation.

BERTScore Evaluates two texts based the cosine similarities between their word embedding representations (Zhang et al., 2019b). Semantically similar texts have many tokens which are co-located in the embedded space, thus pairwise cosine similarities between the texts are high.

MoverScore Similar to BERTScore, MoverScore (Zhao et al., 2019) computes the distance between the embedded candidate and reference summaries. MoverScore does this by solving the constrained Word Mover's Distance optimization problem.

BLEURT A version of BERT (Devlin et al., 2018) pre-trained explicitly to act as an evaluation metric for natural language generation tasks (Sellam et al., 2020). It can additionally be fine-tuned on human ratings, although we choose not to do this to preserve comparability with the other metrics.

3 Datasets

CNN/DailyMail, arXiv and Pubmed are the core summarization datasets used in this study. Summary statistics for these datasets can be found in Table 8 and sample summaries in the supplementary materials. The Quora Question Pairs and annotated CNN/DailyMail datasets are the focus of the evaluation-metric analysis in § 4.1.

CNN/DailyMail Contains articles from the DailyMail and CNN newspapers paired with summaries (in the form of story highlights) written by the same author, which act as the ground truth (Hermann et al., 2015).

arXiv & PubMed Two long document summarization datasets (Cohan et al., 2018). The task is to reproduce the article abstract, which operates as the ground-truth summary. `arXiv.org` and `PubMed.com` are online repositories containing scientific research papers, primarily from maths, computer science and engineering for the former and biomedical and life sciences for the latter.

Quora Question Pairs The Quora Question Pairs (QQP) dataset (Sharma et al., 2019b) tests a system's Natural Language Understanding (NLU)[2]. QQP is composed of 404K pairs of questions published by users to the question answering forum, `www.quora.com` (examples in Table 19). The objective is binary classification to determine if two questions are duplicates.

Annotated CNN/DailyMail Allows measuring the correlation between human scores and automatic metrics' scores for summaries. This contains 500 samples from the CNN/DailyMail dataset. Chaganty et al. (2018) produce four summaries for each sample using text summarization models and tasked human evaluators with scoring each summary based on fluency, focus and overall quality. Additionally, each sample has the reference summary.

4 Methodology

4.1 Evaluation Metrics

Metric	Tokenizer	Type	Vocab size
BERTScore	RoBERTa	BPE	50k
BARTScore	BART	BPE	50k
MoverScore	BERT	WordPiece	30k
BLEURT	BERT	WordPiece	30k

Table 1: The pre-trained tokenizer used by each of the evaluation metrics.

Following Zhang et al. (2019a), we use the Google Research package for ROUGE preprocessing[3], performing tokenization and stemming but not stopword removal. The model-based metrics use transformers which have paired tokenizers for the preprocessing pipeline (outlined in Table 1).

It is important to note that we do not train any part of these metrics. We feel this is critical as we would like a universally applicable metric "off the shelf", in the same vein as ROUGE. However, this presents a potential limitation when it comes to uncommon words, as would be expected in the arXiv and PubMed datasets. Through the use of sub-word tokenization, the metrics are able to compute embeddings for rare words; however, there is no guarantee that these representations will be meaningful. This is potentially why the model-based metrics may not perform well on arXiv and PubMed.

Here we also introduce our novel metric, BARTScore, analogous to BERTScore except for our use of BART (12 encoder and decoder layers, 16 attention heads and 1,024 embedding size,

[2] `https://tinyurl.com/y7co9wuh`

[3] `https://tinyurl.com/yyjtdgy9`

(Lewis et al., 2019)) in place of RoBERTa (Liu et al., 2019). We conjecture that BART's pretraining focus on generating spans means that it forms better representations of longer sequences, thus aiding automatic evaluation.

4.1.1 Human-Metric Correlation

We can compare the relative performances of the evaluation metrics by their relative correlations with human judgement. Inspired by Böhm et al. (2019), we use the annotated CNN/DailyMail dataset introduced in § 3 and compute the correlation between human judgement scores and the evaluation metric scores for each article, giving the results displayed in § 5.1.1.

4.1.2 Quora-Question Pairs

We would like our evaluation metrics to grasp the semantic similarity between two texts and this is a question of NLU, which we can test using the QQP dataset (Sharma et al., 2019b). This dataset contains pairs of (often similarly worded) questions and the task is to identify if these questions are semantically equivalent. Results are shown in § 5.1.2.

4.1.3 Adversarial Analysis

To further probe the effectiveness of the metrics, we performed a set of adversarial tests. These consist of corrupting a set of summaries and assessing how well the evaluation metrics can distinguish the un-corrupted from the corrupted summaries. As PEGASUS's summaries are not significantly worse than human quality (Zhang et al., 2019a), we used these as a proxy for a second set of reference summaries. These experiments used the CNN/DailyMail and PubMed datasets and results are shown in § 5.1.3.

Corruption Methods Our chosen methods of corruption were BERT mask-filling, word-dropping and word permutation, inspired by (Sellam et al., 2020). For each of these methods, the input summary was tokenized and chunked into sequences of length $w \in N$ and the corruption was performed once to each of these sequences. This method ensured that the corruption spans across sentences, thereby gauging the sensitivity of the metric to coherence and grammaticality. By varying the chunk size w we can also determine the sensitivity of the metrics to varying levels of corruption.

BERT mask-filling is a denoising auto-encoding task whereby some of the input tokens are masked and a pre-trained BERT is used to predict these. Word-dropping corrupts the summary by omitting tokens, mimicking some of the common "pathological" issues encountered with automatic summarizers (Sellam et al., 2020). Word-permutation switches the ordering of two adjacent tokens throughout the summary, testing sensitivity to syntax.

4.2 Long Document Summarization

4.2.1 LED vs BART

This section tests whether recent evidence that approximate and regular self-attention perform similarly (e.g. (Kitaev et al., 2020)) holds for text summarization. We modify BART fine-tuned on the CNN/DailyMail dataset to create the 1,024-width LED using the procedure outlined in § 2.1, with an attention window of 512 tokens (most comparable to BART's self-attention). Full model configurations are detailed in Table 6. We fine-tune both models for two epochs on the PubMed dataset and for one epoch on the arXiv and CNN/DailyMail datasets. Hyperparameters were selected using limited grid-search over salient features and were used for subsequent experiments. Results are shown in § 5.2.1.

We are also interested in the performance of different configurations of the LED. There are two hypotheses to test here: 1) using a "longer" model is beneficial for summarizing longer documents; 2) reducing the attention window size will moderately reduce performance, but not catastrophically so. To this end, we experiment with varying the model size and window size using the LED (results in § 5.2.2).

4.2.2 Random Starts Analysis

Intuitively, it seems that using a model with a longer context window would perform better on long document summarization tasks. However, this may not be the case for two reasons: 1) if the salient information is clustered towards the beginning of the document there is no advantage to using a longer input. Given the arXiv/PubMed samples are academic documents, this seems plausible. 2) Summarizing longer documents is innately more challenging as it requires more information distillation.

We introduce a novel *Random Starts* task to evaluate these two hypotheses. Here the input document is truncated both at the beginning and end rather than solely at the end. Now, any sequence of L adjacent tokens can be used as the model input, with the starting index S drawn from a random uniform distribution $S \sim U(0, N - L)$ (document length N and model length L). This is a crude but effective method of ensuring that longer models have a higher probability of seeing the salient information than shorter models. Algorithm 1 outlines the process for implementing this.

Algorithm 1:
Random Starts Truncation

```
Result: Set of truncated documents, T = $\{t_1, \ldots, t_n\}$
for source document $d_i \in D$ do
    Tokens to truncate
      $m_i = length(d_i) - model\ length\ L$
    Draw starting position $s_i,\ s_i \sim U(0, m_i)$
    $t_i = d_i[s_i : s_i + L]$
end
```

5 Results

5.1 Evaluation Metric Experiments

5.1.1 Human-Metric Correlations

Metric	ρ	r	τ
BARTScore	0.308	0.335	0.232
BERTScore	0.307	0.340	0.231
Mover-2	0.253	0.272	0.190
Mover-1	0.243	0.262	0.183
BLEURT	0.240	0.249	0.181
ROUGE-1	0.173	0.192	0.129
ROUGE-2	0.135	0.126	0.101
ROUGE-L	0.127	0.131	0.095

Table 2: Performance on the annotated CNN/DailyMail correlation task by evaluation metric. Spearman ρ, Pearson r and Kendall τ correlations are displayed.

The results for the correlation experiments are displayed in Table 2. As expected, the ROUGE metrics correlate poorly with human evaluator scores and perform worse than all of the model-based metrics. Of the model-based metrics, BARTScore and BERTScore perform best, with BLEURT and MoverScore clustered around the mid-point. Significance tests for differences in metric Pearson correlation can be performed using the William's test (Enderlein, 1961), using the methodology from (Graham, 2015). These show that all model-based metrics are significantly more correlated with human judgement than ROUGE. BARTScore and BERTScore are significantly more correlated than all other metrics at $\alpha = 0.01$, but there is no significant difference between them. These results support our hypothesis that the model-based metrics better reflect semantics than ROUGE.

5.1.2 Quora-Question Pairs

Table 3 displays the results of the QQP task. The figures displayed are the performance of a binary classifier using only the metric score as input. To compute this, we split the data equally into train/test and learned a decision boundary to predict the test set. The results support our hypothesis that contextualized embeddings are beneficial as semantic similarity is easier to determine in the embedded space. Of the model-based metrics, BLEURT performs best with the others clustered around the mid-point.

Metric	Binary Classification	
	Acc.	F1
BLEURT	**0.725**	**0.617**
Mover-2	0.690	0.542
Mover-1	0.687	0.532
BERTScore	0.688	0.546
BARTScore	0.680	0.525
ROUGE-L	0.651	0.471
ROUGE-1	0.651	0.459
ROUGE-2	0.630	0.338

Table 3: Performance on the QQP binary classification task. Shown is the performance of a decision boundary using only the metric output as the solitary feature; accuracy and F1 score.

5.1.3 Adversarial Analysis

Table 4 contains the results for the three adversarial tasks described in § 4.1.3. Figure 1 displays the results when we vary the corruption chunk size, with a shallower slope implying the metric was more sensitive to corruption. BERTScore and BARTScore are evidently the best-performing metrics on these tasks. Using a two-sample T-Test for difference in means and $\alpha = 0.01$, BARTScore is significantly the best on Word-Dropping for CNN/DM and Mask-Filling on both datasets, while there is no significant difference between BERTScore and BARTScore for the other tests. Additionally, the model-based metrics mostly outperform ROUGE across all tasks and datasets, corroborating our findings from § 5.1.1. Results suggest the model-based metrics are robust to common pathologies such as syntactical and lexical discrepancies.

One concern with the model-based metrics is that they might perform poorly on the PubMed dataset because their LMs were not exposed to

Metric	W-D		M-F		W-P	
	C/D	PM	C/D	PM	C/D	PM
BA	**94.6**	**95.5**	**98.0**	**98.2**	**97.6**	**97.4**
BE	92.9	**94.8**	95.4	95.7	**97.8**	**97.9**
M1	86.2	88.6	84.7	88.3	91.9	93.1
M2	83.2	84.6	82.1	85.5	87.7	89.5
BL	70.4	49.8	82.2	86.0	92.8	92.4
R1	78.2	78.6	73.5	89.6	00.0	00.0
R2	74.4	87.8	65.7	88.8	78.5	90.2
RL	77.0	77.3	71.4	85.4	53.8	57.0

Table 4: Mean accuracy on the corruption tasks (word-dropping, BERT mask-filling and word-permutation) on the CNN/DailyMail and Pubmed datasets. Scores within 1% of the maximum score are bold. All standard deviations were small (less than 0.3%).

the biomedical lexicon during pre-training. This would prevent the models from forming meaningful representations of these tokens; however, this does not manifest here as the model-based metrics' performance stays roughly constant across the two datasets.

Also of note is BLEURT's poor performance, particularly on word-dropping and mask-filling. This is surprising as these methods were inspired by those used to generate BLEURT's synthetic pre-training corpus and was therefore expected to excel on these tasks.

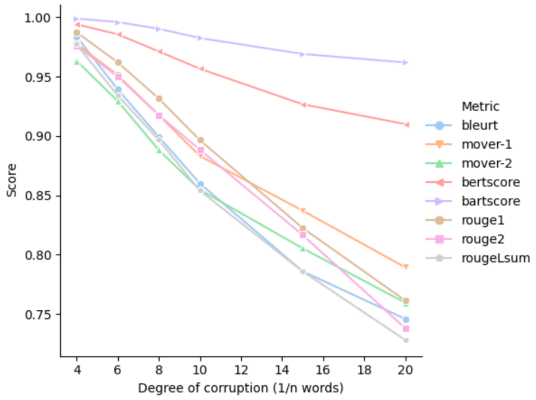

Figure 1: Performance of the metrics on the adversarial tasks as the degree of corruption is varied. The x-axis shows the amount of corruption (1/value) and the y-axis shows the accuracy.

5.2 Long Document Summarization

5.2.1 LED vs BART

§ 2.1 outlined our hypothesis that TED models will perform similarly when using dense and approximate self-attention. This section contains the results of our experiments benchmarking the LED against BART. The results comparing LED-1024 and BART in Table 5 affirm our hypothesis as the LED never performs significantly worse than BART. This is unsurprising: n^2 self-attention and sliding window self-attention are similar when using a 512 window size with 1024 input size[4]. Sliding window self-attention does not convolve (i.e. the first token only attends to the first 512 tokens; likewise the final token only attends to the final 512 tokens). The 513^{th} token attends to all tokens in the sequence. Hence BART and the LED are similar when using 1024-length inputs; it is only when we increase the input length further that the models diverge[5].

5.2.2 LED Performance

§ 4.2.1 outlined our methodology for assessing the performance of the LED. Here we examine the impact of changing the input length and the attention window size on model performance. There are two main results here: 1) using a longer version of the LED improves performance on the arXiv dataset but not using PubMed; 2) an 8x reduction in the attention window size only reduces the performance of the LED by 3.5% on average.

Performance by input length Results are available in Figure 2 and Table 9. Figure 2 shows a side-by-side comparison of input-length against (normalized) metric scores when using Beginning Starts vs Random Starts (see § 4.2.2). Here we are analyzing the normal case corresponding to the first and third figures of Figure 2. This shows there is at best a modest improvement in model performance when using longer model configurations on the PubMed dataset. In § 4.2.2 we hypothesized that lengthening a model may not improve performance if the salient information is clustered at the beginning of the document. Figure 2 supports this for PubMed.

In contrast, there is a clearer positive trend using arXiv. This shows that longer versions of the LED can outperform shorter versions, suggesting the benefits of using a longer context window can outweigh the increased information distillation challenges from using longer inputs. Note also that the LED performed particularly well on the arXiv dataset, competitive with the SOTA (Zaheer et al.,

[4] Recall that attention window is double-sided.

[5] we were unable to reproduce the results from (Lewis et al., 2019) due to differences in the implementation libraries, explaining why our best-performing configuration does not match the author's. This is discussed here: https://tinyurl.com/y4s8jkcd.

	Model	BARTScore	BERTScore	Mover1	Mover2	BLEURT	Rouge2
[1]	BART	0.597 ±0.08	0.302 ±0.13	0.203 ±0.14	0.268 ±0.12	-0.249 ±0.12	0.204 ±0.13
	LED	0.599 ±0.08	0.303 ±0.13	0.203 ±0.14	0.269 ±0.12	-0.237 ±0.27	0.201 ±0.13
[2]	BART	0.606 ±0.01	0.275 ±0.11	0.181 ±0.11	0.237 ±0.10	-0.047 ±0.17	0.188 ±0.13
	LED	0.603 ±0.01	0.270 ±0.11	0.172 ±0.11	0.229 ±0.10	-0.053 ±0.17	0.184 ±0.12
[3]	BART	0.597 ±0.04	0.268 ±0.07	0.161 ±0.08	0.217 ±0.07	-0.091 ±0.15	0.166 ±0.07
	LED	0.597 ±0.04	0.265 ±0.08	0.156 ±0.09	0.213 ±0.08	-0.097 ±0.16	0.165 ±0.08

Table 5: Mean scores and standard deviations for BART and LED-1024 on the CNN/DailyMail [1], PubMed [2] and arXiv [3] datasets. ROUGE-1/L excluded for readability.

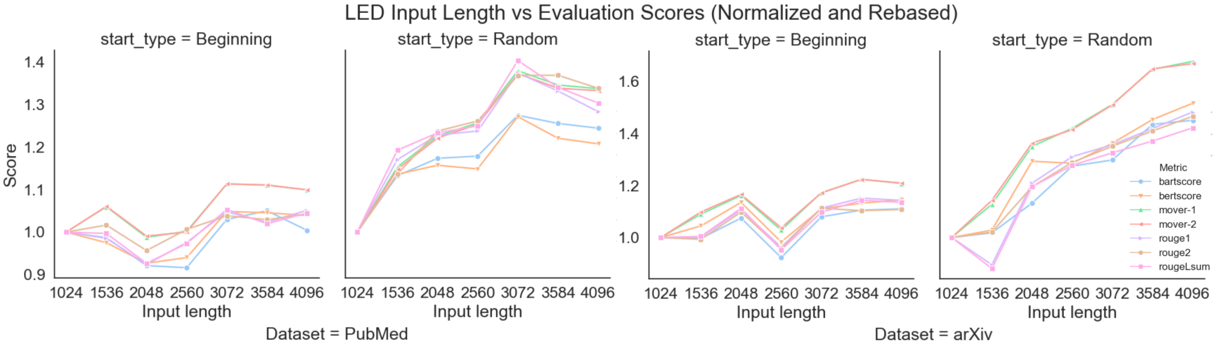

Figure 2: Performance of the LED by model length on the arXiv and PubMed summarization tasks. The left-sided plots show the Beginning Starts case and the Random Starts in the right-sided plots. These results are normalized for better visibility. The raw version of these results are in tables 9 and 10.

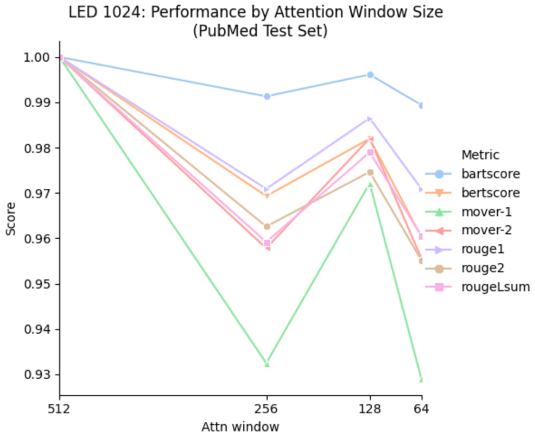

Figure 3: Performance of the `LED-1024` on the PubMed summarization task by attention window length. Un-normalized results are in Table 7.

2020).

Compare this now to the Random Starts case, displayed in the second and fourth figures in Figure 2. Here we see that performance increases dramatically with model length, corroborating our hypothesis from § 4.2.1 that the salient content is clustered towards the beginning of the document and therefore the datasets are biased toward shorter models.

Performance by attention window The results (Figure 3) show that reducing the attention window size has a marginal negative impact on the LED's performance. Cutting the LED-1024's attention window from 512 to 64 resulted in a 1% - 7% fall in performance. These results suggest that self-attention can be made more efficient with limited performance degradation. This is significant as it greatly reduces the memory consumption and cost of long document summarization. For example, with an attention window of 512 and a batch size of 1, the maximum sequence length that can fit on a 12 Gb GPU is 1,024. This rises to 2,560 if the window size is 64.

6 Discussion

Qualitative Analysis of Summaries The supplementary materials contain samples from each dataset with their corresponding target and generated model summaries. These show that the models produce coherent and relevant summaries for all datasets and adjust well to the diverse discourse styles. Previous studies ((Zhang et al., 2019a), (Lewis et al., 2019)) performed human experiments using the CNN/DailyMail dataset and concluded their output summaries are not significantly worse than the references. We judge the LED's summaries to be of similarly high standard.

There is, however, a quality gap between the human abstracts and the model summaries on PubMed and arXiv. It is still an open research question of how to verify the factual authenticity of generated summaries with erroneous claims being a widespread problem. These issues are common for all models we experimented with on arXiv and PubMed. We conjecture this is partly due to the lack of overlap between the lexicon used in these articles with the LMs' pre-training corpora.

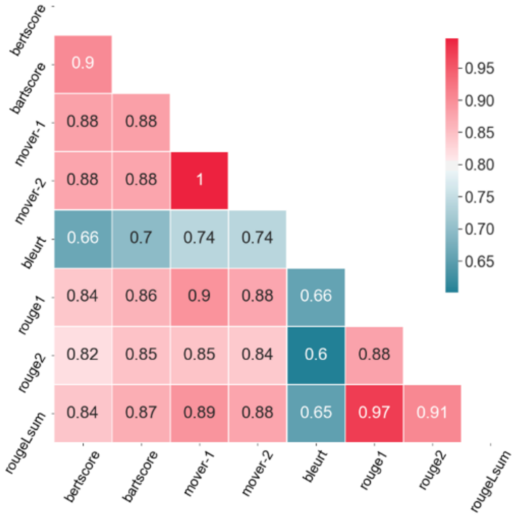

Figure 4: Correlation matrix (Pearson r) of evaluation metrics' scores using the summaries produced by ProphetNet, BART and PEGASUS on the CNN/DailyMail test set.

Comparison of Metrics It was a recurring theme in § 5.1 for BLEURT to behave erratically. It had the strongest performance on the QQP task from § 5.1.2 but was poor on the adversarial tests in § 5.1.3. Figure 4 displays the correlation matrix for metric scores on CNN/DailyMail summaries and this shows the metrics are generally highly correlated. Following (Graham, 2015), we use the William's test (Enderlein, 1961) for significant differences in the correlations between metrics. We find that BLEURT's correlation with each metric is significantly lower than for every other metric. Coupled with inconsistent performance, this raises a red flag, hence we recommend against using BLEURT for summarization. We believe BARTScore and Mover-2 are best as they are more correlated with human judgement than BERTScore and Mover-1 and are less correlated with each other. However, even BARTScore, the best-performing metric, has low correlation with human judgement, indicating there is some way to go before these metrics can substitute for human evaluation for summarization.

7 Concluding Remarks

Future work Several self-attention approximations have been created in addition to Longformer (e.g. (Wang et al., 2020), (Qiu et al., 2019), (Child et al., 2019)). Future work could systematically benchmark the performance of TEDs constructed using these layers. Alternatively, one could add an additional pre-training step so the LED-4096 is pre-trained in the same fashion as BART or PEGASUS. Our experiments could also be repeated on the BIGPATENT (Sharma et al., 2019a) dataset as the salient content is uniformly distributed throughout the source documents here.

Conclusion This study has highlighted evaluation and long document summarization as bottlenecks for the field of text summarization. Here we have shown that a set of model-based evaluation metrics outperform ROUGE over a wide range of tasks. We have also shown that sparse self-attention is an effective method of reducing the memory complexity of transformers which can therefore be more effectively applied to longer documents.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. pages 2

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. *arXiv preprint arXiv:1909.01214*. pages 2, 4

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics. pages 3

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509. pages 8

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *CoRR*, abs/1804.05685. pages 1, 3, 12

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. pages 2, 3

G. Enderlein. 1961. Williams, e. j.: Regression analysis. wiley, new york 1959, 214 s., $ 7,50. *Biometrische Zeitschrift*, 3(2):145–145. pages 5, 8

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model. *CoRR*, abs/1906.01749. pages 1

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China. Association for Computational Linguistics. pages 5, 8

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340. pages 1, 3

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. pages 2, 4

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. pages 2, 4, 6, 7

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. pages 1, 2

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. pages 4

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023. pages 1

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. pages 1

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. pages 3

Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen tau Yih, Sinong Wang, and Jie Tang. 2019. Blockwise self-attention for long document understanding. pages 8

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning Representations by Back-propagating Errors. *Nature*, 323(6088):533–536. pages 1

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 1

Horacio Saggion and Thierry Poibeau. 2013. *Automatic Text Summarization: Past, Present and Future*, pages 3–21. Springer Berlin Heidelberg, Berlin, Heidelberg. pages 1

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. pages 1, 12

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. pages 3, 4

Eva Sharma, Chen Li, and Lu Wang. 2019a. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics. pages 8

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019b. Natural language understanding with the quora question pairs dataset. *CoRR*, abs/1907.01041. pages 3, 4

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215. pages 1

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762. pages 1, 2

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. pages 2, 8

R. J. Williams and D. Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280. pages 2

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. pages 2

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. pages 2, 6

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. pages 2, 3, 4, 7

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675. pages 3

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. pages 3

Appendices

A Supplementary Materials

Parameter	**BART-LG**	**LED**
Attention type	n^2	Window-512
Max. enc. input len.		
- CNN/DM	1024	1024*
- PubMed	1024	4096
- arXiv	1024	4096
Max. dec. input len.		
- CNN/DM	142	142
- PubMed	200	200
- arXiv	200	200
Beam size	4	4
Length penalty	2.0	2.0
Num. of heads	16	16
Num. layers	12	12
Hidden layer dim.	1024	1024
Batch size	1	1
Activation function	GeLU	GeLU
Optimizer	AdamW	AdamW
Learning rate	3e-5	1e-5
Label smoothing	0.1	0.0
Dropout	0.1	0.1
Chunk size (local)	N.A.	N.A.
Chunk size (LSH)	N.A.	N.A.

Table 6: Model configurations of our best-performing variants of each model after performing hyperparameter search. * We did not use a longer version of the `LED` for the CNN/DailyMail dataset as the articles are short.

Attn Window	BA	BE	M-1	M-2	BLEURT	R-1	R-2	R-L
512	0.599	0.264	0.167	0.224	-0.055	0.428	0.176	0.383
	0.06	0.11	0.11	0.10	0.17	0.00	0.00	0.00
256	0.594	0.256	0.156	0.215	-0.078	0.416	0.170	0.367
	0.06	0.11	0.11	0.10	0.18	0.00	0.00	0.00
128	0.597	0.260	0.163	0.220	-0.066	0.423	0.172	0.375
	0.06	0.11	0.11	0.10	0.17	0.00	0.00	0.00
64	0.593	0.254	0.155	0.214	-0.080	0.416	0.168	0.368
	0.06	0.11	0.11	0.10	0.17	0.00	0.00	0.00

Table 7: LED-1024 performance on a summarization task by attention window size. Scores reported are the mean metric scores with the standard deviation underneath. These scores are obtained using the PubMed test set after fine-tuning the models for one epoch on the PubMed dataset. These results are displayed graphically in Figure 3.

Dataset	# docs (K)	Avg. doc # words	Avg. sum # words	% over 1,024 toks	% over 4,096 toks	Citations	Year
CNN	92	656	43	8%*	0%	998	2017
DailyMail	219	693	52	8%*	0%	998	2017
PubMed	133	3,016	203	87%	24%	81	2019
arXiv	215	4,938	220	97%	61%	81	2019

Table 8: Descriptive statistics of the datasets used in this study. The fifth and sixth columns indicate the share of each dataset that would not fit in BART and LED-4096 models given their capacities of 1,024 and 4,096 tokens respectively. The number of citations is the Google Scholar citations from the original papers ((Cohan et al., 2018), (See et al., 2017)) as of 24/08/2020. * The CNN/DailyMail dataset is 8% overall.

Model Length	BA	BE	M-1	M-2	BLEURT	R-1	R-2	R-L
PubMed								
1024	0.596	0.262	0.159	0.217	-0.068	0.423	0.174	0.375
	0.06	0.11	0.11	0.10	0.17	0.10	0.11	0.10
1536	0.595	0.259	0.165	0.224	-0.071	0.422	0.176	0.375
	0.06	0.11	0.11	0.10	0.18	0.10	0.12	0.11
2048	0.592	0.254	0.157	0.217	-0.080	0.416	0.169	0.367
	0.06	0.11	0.11	0.10	0.18	0.10	0.12	0.11
2560	0.591	0.256	0.159	0.217	-0.070	0.421	0.174	0.372
	0.06	0.11	0.11	0.10	0.18	0.10	0.12	0.11
3072	0.598	0.268	0.171	0.229	-0.054	0.428	0.178	0.380
	0.06	0.10	0.10	0.10	0.17	0.10	0.11	0.10
3584	0.599	0.267	0.171	0.229	-0.061	0.426	0.177	0.377
	0.06	0.11	0.11	0.10	0.17	0.10	0.12	0.10
4096	0.596	0.267	0.170	0.227	-0.060	0.429	0.179	0.380
	0.06	0.11	0.11	0.10	0.17	0.10	0.12	0.10
arXiv								
1024	0.597	0.265	0.156	0.213	-0.097	0.439	0.165	0.388
	0.04	0.08	0.09	0.08	0.16	0.08	0.08	0.08
1536	0.597	0.269	0.164	0.221	-0.094	0.439	0.165	0.389
	0.04	0.08	0.08	0.08	0.16	0.08	0.08	0.08
2048	0.601	0.276	0.170	0.226	-0.080	0.448	0.173	0.397
	0.04	0.08	0.09	0.08	0.16	0.08	0.08	0.08
2560	0.594	0.264	0.158	0.216	-0.098	0.436	0.162	0.384
	0.04	0.07	0.08	0.07	0.16	0.08	0.08	0.08
3072	0.601	0.274	0.171	0.227	-0.084	0.448	0.174	0.396
	0.04	0.08	0.09	0.08	0.16	0.08	0.08	0.08
3584	0.602	0.276	0.175	0.231	-0.074	0.451	0.174	0.400
	0.04	0.08	0.08	0.08	0.16	0.08	0.08	0.08
4096	0.602	0.277	0.174	0.230	-0.073	0.451	0.174	0.399
	0.04	0.08	0.09	0.08	0.16	0.08	0.08	0.08

Table 9: LED performance on the summarization task by model length with a 512 attention window (using Beginning Starts, PubMed and arXiv test sets). Scores reported are the mean metric score with the standard deviation underneath. These results are displayed graphically in Figure 2 (left-sided panels).

Model Length	**BA**	**BE**	**M-1**	**M-2**	**BLEURT**	**R-1**	**R-2**	**R-L**
PubMed								
1024	0.577 0.05	0.231 0.09	0.121 0.09	0.184 0.08	-0.105 0.16	0.385 0.09	0.128 0.08	0.336 0.08
1536	0.584 0.05	0.245 0.09	0.136 0.09	0.197 0.08	-0.087 0.16	0.402 0.09	0.143 0.09	0.355 0.09
2048	0.586 0.06	0.247 0.10	0.144 0.10	0.204 0.09	-0.085 0.17	0.407 0.10	0.153 0.11	0.359 0.10
2560	0.587 0.06	0.246 0.11	0.147 0.10	0.207 0.10	-0.083 0.17	0.408 0.10	0.156 0.11	0.361 0.10
3072	0.592 0.05	0.259 0.10	0.159 0.10	0.218 0.09	0.421 0.09	0.167 0.11	0.375 0.09	-0.061 0.16
3584	0.591 0.06	0.254 0.11	0.156 0.11	0.215 0.10	-0.076 0.17	0.417 0.10	0.167 0.12	0.369 0.10
4096	0.592 0.05	0.259 0.10	0.159 0.10	0.218 0.09	-0.061 0.16	0.421 0.09	0.167 0.11	0.375 0.09
arXiv								
1024	0.574 0.04	0.226 0.07	0.093 0.08	0.155 0.08	-0.168 0.17	0.395 0.08	0.126 0.06	0.351 0.07
1536	0.575 0.05	0.228 0.08	0.104 0.09	0.167 0.08	-0.173 0.18	0.386 0.09	0.127 0.07	0.341 0.08
2048	0.593 0.04	0.260 0.08	0.150 0.09	0.208 0.08	-0.103 0.16	0.431 0.08	0.155 0.08	0.380 0.08
2560	0.580 0.04	0.248 0.07	0.124 0.08	0.185 0.08	-0.128 0.16	0.413 0.08	0.140 0.07	0.366 0.08
3072	0.586 0.04	0.248 0.07	0.130 0.08	0.189 0.08	-0.126 0.16	0.422 0.08	0.147 0.07	0.373 0.08
3584	0.587 0.04	0.254 0.07	0.138 0.08	0.197 0.08	-0.120 0.16	0.426 0.08	0.151 0.07	0.377 0.08
4096	0.594 0.04	0.265 0.08	0.152 0.09	0.210 0.08	-0.100 0.16	0.436 0.08	0.160 0.08	0.384 0.08

Table 10: LED performance on a summarization task by model length. This is for the arXiv and PubMed test sets using Random Starts (see section 4.2.2 for details). Scores reported are the mean metric score with the standard deviation underneath. These results are displayed graphically in Figure 2 (right-sided panels).

Highest-scored summaries	
Targets	**Generations**
More details of the so-called ’everyday Americans’ have been revealed. Gardener Julie Stauch was state campaign manager for Wendy Davis. Sean Bagniewski, who appears with his wife, has campaigned for Clinton. But he held an event for likely rival O’Malley days before video’s release.	Julie Stauch, Sean Bagniewski and Vidhya Reddy all appear in Clinton’s video. They are part of a diverse group of ’everyday Americans’ chosen to show aspirational citizens as Clinton looks to speak to a wide demographic of voters. But far from being ’ordinary Americans’, they are directly involved with the Democrat party - just like Jared Milrad and Nate Johnson.
Ellanora Arthur Baidoo has been trying to divorce her husband for several years. Husband doesn’t have permanent address or permanent employment. Baidoo is granted permission to send divorce papers via Facebook.	Ellanora Arthur Baidoo has been trying to divorce her husband for several years, her lawyer says. A New York judge says Facebook is an acceptable way for her to serve her husband with a divorce summons. The couple never lived together, according to court documents.
Michael Gridley, 26, was jailed after running the scam at store in Basildon. Was sacked from position after supermarket received anonymous reports. But he is now employed as a manager at Lidl supermarket in Romford. Sentenced to 12 months at Southend Crown Court for leading role in scam.	Michael Gridley, 26, was jailed for a year at Southend Crown Court. He stole £15,000 worth of goods from Asda in Basildon, Essex. Stock including alcohol, cigarettes and DVDs were taken from the store. Gridley was sacked

Scores							
0.615	0.879	0.891	0.894	0.946	0.950	0.949	0.950
0.748	0.823	0.839	0.895	0.936	0.935	0.933	0.935
0.596	0.838	0.853	0.864	0.895	0.896	0.723	0.896

Table 11: Three examples of highly scored summaries. These were rated in the top 1% by BARTScore, Mover-1, ROUGE-1 and ROUGE-L. Summaries from the CNN/DailyMail test set and produced by the LED-1024. The evaluation scores for each of the summaries is shown below; these correspond to BLEURT, Mover-1, Mover-2, BERTScore, BARTScore, ROUGE-1, ROUGE-2 and ROUGE-L (in order).

Lowest-scored summaries	
Targets	**Generations**
Kenya’s security has been bogged down by concerns over civil rights. Kenyan Muslims have been targeted in raids and robbed, says Human Rights Watch.	Al-Shabaab killed 147 people at a college campus in Garissa, Kenya, on Thursday. The number of people killed is plaguing Kenyans with self-doubt, CNN’s David McKenzie says. Kenya’s politicians and public have struggled with these ideas...
Indiana town’s Memories Pizza is shut down after online threat. Its owners say they’d refuse to cater a same-sex couple’s wedding.	Memories Pizza in Indiana is at the center of the debate over the state’s Religious Freedom Restoration Act. "If a gay couple was to come and they wanted us to bring pizzas to their wedding, we’d have to say no," owner says. Critics say the law
Mohonk Mountain House is a ’castle’ retreat 90 minutes from New York. The hotel sits blissfully on the banks of Lake Mohonk in the Hudson Valley. The hotel was originally built as a drinking inn 145 years ago before Quaker twins Albert and Alfred Smiley made it a dry retreat - the bar is now open.	Mohonk Mountain House is a faux-gothic Victorian castle in the heart of the Hudson Valley. The lake, gardens and trails are a vast adventure playground for all ages. The 360-degree views are inspirational and the kids’ club is the best we have

Scores							
-0.903	-0.142	-0.051	0.011	0.397	0.051	0.000	0.051
-0.699	-0.083	0.003	0.016	0.436	0.136	0.023	0.091
-0.901	-0.214	-0.102	0.036	0.358	0.070	0.000	0.070

Table 12: Three examples of poorly scored summaries. These were rated in the bottom 1% by BARTScore, Mover-1, ROUGE-1 and ROUGE-L. Summaries from the CNN/DailyMail test set and produced by the LED-1024. The evaluation scores for each of the summaries is shown below; these correspond to BLEURT, Mover-1, Mover-2, BERTScore, BARTScore, ROUGE-1, ROUGE-2 and ROUGE-L (in order).

<table>
<tr><th colspan="8">Highly scored by model-based metrics, poorly scored by ROUGE</th></tr>
<tr><th colspan="4">Targets</th><th colspan="4">Generations</th></tr>
<tr><td colspan="4">Liverpool scouts have been impressed by Geoffrey Kondogbia this season. The midfielder was one of the most coveted youngsters in Europe. France international joined Monaco from Sevilla in 2013 for £17million. Liverpool remain in the frame for James Milner and Danny Ings.</td><td colspan="4">Liverpool are watching Monaco midfielder Geoffrey Kondogbia. France international has impressed in Europe and Ligue 1 this season. Real Madrid, Manchester United, Juventus and PSG were all keen. Brendan Rodgers’ side are also interested in Danny Ings and James Milner.</td></tr>
<tr><td colspan="4">Jeremy Trentelman, 36, of Ogden, built fort for young son and daughter. He received letter one day later saying it violated ordinance against waste. Father plans on keeping castle up for 14 days before he receives fine.</td><td colspan="4">Jeremy Trentelman, 36, of Ogden, Utah, last week built a giant box fort for his son Max, 3, and daughter Story, 2, that included trap doors and a small slide. The father, who works as a florist arranging intricate displays...</td></tr>
<tr><td colspan="4">Sir Bradley Wiggins left Team Sky after Paris-Roubaix on April 12. Tour de Yorkshire begins in Bridlington and finishes in Leeds from May 1-3. Wiggins’ eponymous team is completed by Steven Burke, Mark Christian, Andy Tennant, Owain Doull and Jon Dibben.</td><td colspan="4">Bradley Wiggins will ride for his eponymous team in the Tour de Yorkshire. The 2012 Tour de France winner was not selected in Team Sky’s 2014 squad. The Tour begins in Bridlington and finishes in Leeds on May 3. It is a legacy of the Grand Depart</td></tr>
<tr><th colspan="8">Scores</th></tr>
<tr><td>-0.022</td><td>0.285</td><td>0.346</td><td>0.174</td><td>0.634</td><td>0.356</td><td>0.169</td><td>0.329</td></tr>
<tr><td>0.057</td><td>0.275</td><td>0.337</td><td>0.435</td><td>0.644</td><td>0.343</td><td>0.206</td><td>0.323</td></tr>
<tr><td>0.006</td><td>0.372</td><td>0.446</td><td>0.398</td><td>0.649</td><td>0.327</td><td>0.104</td><td>0.327</td></tr>
</table>

Table 13: Three examples of summaries scored well by the model-based metrics but scored poorly by the ROUGE metrics. These were rated in the top third of all summaries by BARTScore and Mover-1 but in the bottom third by ROUGE-1 and ROUGE-L. Summaries from the CNN/DailyMail test set and produced by the LED-1024. The evaluation scores for each of the summaries is shown below; these correspond to BLEURT, Mover-1, Mover-2, BERTScore, BARTScore, ROUGE-1, ROUGE-2 and ROUGE-L (in order).

<table>
<tr><th colspan="8">Poorly scored by model-based metrics, highly scored by ROUGE</th></tr>
<tr><th colspan="4">Targets</th><th colspan="4">Generations</th></tr>
<tr><td colspan="4">Father-of-three Craig Sytsma was mauled to death in Michigan last year. Dog owners Sebastiano Quagliata and wife Valbona Lucaj agreed to plea deal Friday to around 15 years in jail, though judge could add six months. Sytsma was jogging in July 2014 when the two cane corsos attacked him. He was bitten almost ten times and was’screaming and begging’ for help.</td><td colspan="4">Sebastiano Quagliata and wife, Valbona Lucaj, pleaded no contest to owning a dangerous dog causing death in the fatal mauling last summer of Craig Sytsma of Livonia, Michigan. ’I’m so sorry,’ Lucaj told Sy...</td></tr>
<tr><td colspan="4">YouTube user Serpentor filmed his feline friend in action. Footage shows the tabby producing bizarre noises as she is petted.</td><td colspan="4">YouTube user Serpentor filmed his feline friend in action. Footage shows tabby producing a range of unusual gurgling noises as she is petted. Her owners are heard laughing in the background as they watch her. To date the clip of her singing has been...</td></tr>
<tr><td colspan="4">Little Catalina from America was filmed emptying out the contents of the kitchen cupboards at home. But when her father tells her to clean up her mess, she vehemently refuses with a heated - and rather cute - argument ensuing.</td><td colspan="4">Catalina from America was filmed emptying out the contents of the kitchen cupboards at home. But when her father tells her to clean up her mess, she vehemently refuses with a heated - and rather cute - argument ensuing. ’I already cleaned the kitchen, no it...</td></tr>
<tr><th colspan="8">Scores</th></tr>
<tr><td>-0.516</td><td>0.143</td><td>0.190</td><td>0.210</td><td>0.541</td><td>0.484</td><td>0.247</td><td>0.462</td></tr>
<tr><td>-0.358</td><td>0.147</td><td>0.188</td><td>0.282</td><td>0.550</td><td>0.533</td><td>0.273</td><td>0.511</td></tr>
<tr><td>-0.510</td><td>0.120</td><td>0.186</td><td>0.199</td><td>0.555</td><td>0.500</td><td>0.184</td><td>0.480</td></tr>
</table>

Table 14: Three examples of summaries scored poorly by the model-based metrics but scored well by the ROUGE metrics. These were rated in the bottom third of all summaries by BARTScore and Mover-1 but in the top third by ROUGE-1 and ROUGE-L. Summaries from the CNN/DailyMail test set and produced by the LED-1024. The evaluation scores for each of the summaries is shown below; these correspond to BLEURT, Mover-1, Mover-2, BERTScore, BARTScore, ROUGE-1, ROUGE-2 and ROUGE-L (in order).

CNN/DailyMail	
Source:	(CNN)Sky watchers in western North America are in for a treat: a nearly five-minute total lunar eclipse this morning. Here's how it's unfolding: . It started at 3:16 a.m. Pacific Daylight Time, when the moon began moving into Earth's shadow. For the next hour and 45 minutes, that shadow will move across the moon and engulf it at 4:58 a.m. Pacific Time. The total eclipse will only last four minutes and 43 seconds, and NASA says that makes it the shortest one of the century. Watch it live on NASA TV . While people west of the Mississippi River will have the best view, at least a partial eclipse will be visible across the nation. But sunrise will interrupt the show on the East Coast. Parts of South America, India, China and Russia also will be able to see the eclipse, but it won't be visible in Greenland, Iceland, Europe, Africa or the Middle East. A lunar eclipse happens when the sun, Earth and moon form a straight line in space, with the Earth smack in the middle. The sun shines on the Earth and creates a shadow. As the moon moves deeper into that shadow, it appears to turn dark and may even appear to be a reddish color. Why red? Because Earth's atmosphere is filtering out most of the blue light. Some people have nicknamed the effect the "blood moon." NASA says lunar eclipses typically happen at least twice a year, but this eclipse is the third in a series of four in a row, known as a "tetrad.". The first was on April 15, 2014. The second was in September 2014, the next is Saturday and there will be one more, on September 28. If you want to learn more about the eclipse, NASA astronomer Mitzi Adams will take questions on Twitter NASA_Marshall. Did you see the total lunar eclipse? Share your photos with CNN iReport.
Target:	The total eclipse will only last 4 minutes and 43 seconds . People west of the Mississippi River will have the best view . Parts of South America, India, China and Russia also will see the eclipse .
LED-1024:	A nearly five-minute total lunar eclipse will be visible in western North America. It will only last four minutes and 43 seconds, and NASA says it's the shortest one of the century. Parts of South America, India, China and Russia will also see the eclipse.
BART:	The total lunar eclipse started at 3:16 a.m. Pacific Daylight Time. People west of the Mississippi River will have the best view. Parts of South America, India, China and Russia also will be able to see the eclipse. It won't be visible in Greenland, Iceland, Europe, Africa or Middle East.
PEGASUS:	A lunar eclipse happens when the sun, Earth and moon form a straight line in space . A lunar eclipse is the third in a series of four in a row, known as a "tetrad"
ProphetNet:	it started at 3 : 16 a . m . pacific daylight time , when the moon began moving into earth 's shadow . for the next hour and 45 minutes , that shadow will move across the moon and engulf it . the total eclipse will only last four minutes and 43 seconds .

Table 15: A sample CNN/DailyMail article with model summaries.

	PubMed
Source:	gossypiboma (retained surgical sponge) describes a mass of cotton or sponge that is left behind in the body cavity during a surgical operation . these foreign bodies can often mimic tumors or abscesses when detected clinically or with radiation . in fact , surgical sponges without opaque markers are the leading cause for the difficulty of a correct diagnosis . here , we present a case in which a foreign body , gauze without radiopaque markers , was left behind during a procedure that involved the removal of an intrapelvic tumor . after a computed tomography (ct) scan had revealed information leading to an inaccurate diagnosis , an accurate diagnosis of gossypiboma was successfully made using endoscopic ultrasound - fine - needle aspiration (eus - fna) . a 60-year - old female patient was referred with an unexpected detection of an intrapelvic tumor during a routine ct scan . the patient 's medical history included a hysterectomy and a left side ovariectomy at the age of 27 . the ct scan revealed a 45 mm tumor in the patient 's pelvis [figure 1a and b] , which was diagnosed as a gastrointestinal stromal tumor (gist) , a chronic expanding hematoma or leiomyoma . a forward - view eus (tgf - uc260j , olympus optical co. , tokyo , japan) identified a low echoic lesion when scanned from the sigmoid colon . fna was performed with a 22-gauge needle (ezshot2 , olympus optical co. , tokyo , japan) by using the forward - view eus [figure 2] . examination of the fna specimen revealed a foreign body (string of gauze) displaying macrophage cells and necrotic tissue [figure 3 ; hematoxylin and eosin , original magnification 40] . during the surgery , histological examination of the tissue adjacent to the foreign body revealed granuloma formations with fragmented silken threads [figure 4 ; hematoxylin and eosin , original magnification , 100 and 200] . a : sagittal section (black arrow) ; b : horizontal section (black arrow) . endoscopic ultrasound (eus)-fine - . . .
Target:	$< S >$ gossypiboma is a term used to describe a foreign mass within the body that is composed of a cotton matrix. $< /S >< S >$ gossypiboma can cause infection or abscess formation from an early stage, or it can remain clinically silent for many years. $< /S >< S >$ gossypiboma can be diagnosed using several methods including abdominal x - ray films, us of the abdomen, contrast - enhanced computed tomography of the abdomen, and during surgery. $< /S >< S >$ however, the variable appearance of gossypiboma can lead to diagnostic misinterpretations. here, we report a successful diagnosis of gossypiboma using a novel method, that is, endoscopic ultrasound - fine - needle aspiration. $< /S >$
LED-4096:	$< S >$ gossypiboma (retained surgical sponge) describes a mass of cotton or sponge that is left behind in the body cavity during a surgical operation. here, we present a case in which a foreign body, gauze without radiopaque markers, was left behind during a procedure that involved the removal of an intrapelvic tumor. after a computed tomography (ct) scan had revealed information leading to an inaccurate diagnosis, an accurate diagnosis of gossiopiboma was successfully made using endoscopic ultrasound - fine - needle aspiration (eus - fna). $< /S >$
BART:	$< S >$ a gossypiboma, also called textiloma, literally means retained surgical sponge and is derived from the latin word " gosspium ", which means " cotton, " and " boma " in kiswahili, " " place of concealment ". " it is difficult to diagnose " textilomas " in the neck " because of its rarity, various symptoms, and non - specific radiologic findings. " " " $< /S >$ " furthermore, it can be misinterpreted as malignancy " and finally diagnosed from a histopathological examination of the surgical specimen after unwarranted radical surgery. " to our knowledge ", " text

Table 16: A sample PubMed article with model summaries. The source has been truncated from 717 to 350 tokens for readability.

arXiv	
Source:	the brain is a computational device for information processing and its flexible and adaptive behaviors emerge from a system of interacting neurons depicting very complex networks @xcite . many biological evidences suggest that the neocortex implements a common set of algorithms to perform " intelligent " behaviors like learning and prediction . in particular , two important related aspects seem to represent the crucial core for learning in biological neural networks : the hierarchical information processing and the abstraction process @xcite . the hierarchical architecture emerges from anatomical considerations and is fundamental for associative learning (e.g. multisensory integration) . the abstraction instead leads the inference of concepts from senses and perceptions (fig . [fig:1]d) . + specifically , information from sensory receptors (eyes , skin , ears , etc .) travels into the human cortical circuits following subsequent abstraction processes . for instance , elementary sound features (e.g. frequency , intensity , etc .) are first processed in the primary stages of human auditory system (choclea) . subsequently sound information gets all the stages of the auditory pathway up to the cortex where higher level features are extracted (fig . [fig:1]e - f) . in this way information passes from raw data to objects , following an abstraction process in a hierarchical layout . thus , biological neural networks perform generalization and association of sensory information . for instance , we can associate sounds , images or other sensory objects that present together as it happens in many natural and experimental settings like during pavlovian conditioning . biological networks process these inputs following a hierarchical order . in a first stations inputs from distinct senses are separately processed accomplishing data abstraction . this process is repeated in each subsequent higher hierarchical layer . doing so , in some hierarchical layer , inputs from several senses converge showing associations among sensory inputs . + recent findings indicate that neurons can perform invariant recognitions of their input activity patterns producing specific modulations of their synaptic releases @xcite . although the comphrension of such neuronal mechanisms is still elusive , these hints can drive the development of algorithms closer to biology than spiking networks or other brain - inspired models appear to be . + in this work , we propose a learning framework based on these . . .
Target:	$< S >$ the human brain processes information showing learning and prediction abilities but the underlying neuronal mechanisms still remain unknown. recently, many studies prove that neuronal networks are able of both generalizations and associations of sensory inputs. $< /S >$ $< S >$ + in this paper, following a set of neurophysiological evidences, we propose a learning framework with a strong biological plausibility that mimics prominent functions of cortical circuitries. $< /S > < S >$ we developed the inductive conceptual network (icn), that is a hierarchical bio - inspired network, able to learn invariant patterns by variable - order markov models implemented in its nodes. $< /S > < S >$ the outputs of the top - most node of icn hierarchy, representing the highest input generalization, allow for automatic classification of inputs. $< /S > < S >$ we found that the icn clusterized mnist images with an error of 5.73% and usps images with an error of 12.

Table 17: A sample arXiv article with model summaries (part 1). The source has been truncated from 4,522 tokens to 400 for readability.

arXiv (continued	
LED-4096:	$< S >$ we propose a general neural learning framework, called inductive conceptual network (icn), where biological neurons are individually identified by nodes and invariant recognition is performed by variable - order markov models (vmm). the icn represents a general biological plausible model of the learning mechanisms in neuronal networks, and we tested the accuracy of this network on the mnist and usps datasets. $< /S >$ < s> we find that the neural network exhibits a complex hierarchy of node nodes, which exhibits invariant pattern recognition by a hidden markov model algorithm implemented into a set of software libraries specialized in image processing (hierarchical temporal memory (htm))).
BART:	$< S >$ the inductive conceptual network (icn) represents a general biological plausible model of the learning mechanisms in neuronal networks. $< /S > < S >$ the icn is based on a set of considerations primarily from the memory - prediction framework, which is a generalization of jeff hawkins _ on intelligence_. we propose a learning framework based on these biological considerations and we tested the accuracy of this network on the mnist and usps datasets. in this framework, we find that neural neurons can perform invariant recognitions of their input activity patterns producing specific modulations of their synaptic releases, which can drive the development of algorithms closer to biology than other brain - inspired models appear to

Table 18: A sample arXiv article with model summaries (part 2).

Quora Question Pairs		
Question 1	Question 2	Equivalent?
Can we ever store energy produced in lightning?	Is it possible to store the energy of lightning?	Yes
What Game of Thrones villain would be the most likely to give you mercy?	What Game of Thrones villain would you most like to be at the mercy of?	Yes
Why do some people think Obama will try to take their guns away?	Has there been a gun control initiative to take away guns people already own?	No
What are the best YouTube channels to learn medicine?	What are some of the best YouTube channels for learning Git?	No

Table 19: Two positive and two negative samples from the QQP dataset. The objective is to ascertain whether the two questions are semantically equivalent and these labels are provided in the final column.