Enhancing the Performance of Abstractive Summarization on Large Text Corpora

Brendan Wang ‘23
Adviser: Danqi Chen

Abstract

Recently, transformer models have achieved high qualitative and quantitative performances in the abstractive summarization of texts. However, existing challenges include the limited input size of models as well as grammatical [12] and factual inconsistencies [2, 15] in generated summaries. Moreover, automatic evaluation metrics used for summarization including the ROUGE score [11] have many limitations [5]. Finally, there is a lack of work comparing the performance and computational costs of different state-of-the-art summarization models. In this work, we study ways to enhance abstractive summarization of the model T5-small [17] using text shortening techniques and compare the performance of T5-small against the much larger state-of-the-art transformer model PEGASUS [22]. Using the Multi-News dataset [4], we fine-tune and evaluate the T5-small and PEGASUS using the ROUGE metric, the BLEURT score [18], and human evaluation. Our results reveal that following text shortening, T5-small outperforms PEGASUS by about 7.5 in ROUGE-2 (an over 10% increase) and over 0.045 in BLEURT (an over 36% increase). Furthermore, the T5-small model demonstrates less memory and time usage for fine-tuning than PEGASUS.

1. Introduction

The digital revolution has drastically increased the amount of information available and accessible to consumers. Given that many information sources are long or detailed, it is important to be able to easily and efficiently retrieve knowledge from such sources. Fortunately, recent advancements in machine learning have made it possible to automatically summarize texts with considerable accuracy, with applications ranging from search queries to news articles to e-commerce [7].

The task of extracting the most important and relevant information from a document and condensing it into a shorter version that still retains the original meaning is known as text summarization [1]. The two main types of text summarization are extractive and abstractive summarization. Extractive summarization refers to generating a summary by combining together the most important elements from the original input. By contrast, abstractive summarization involves generating a novel "paraphrased" summary which includes elements that may have not been present in the original input.

Unlike extractive summarization, abstractive summarization theoretically allows for the generation of human-like summaries that are concise, meaningful, and fluent. In practice, however, this remains a challenge as models still struggle with understanding the input enough to summarize the main points [21]. Indeed, abstractive summarization models still suffer from factual inconsistencies [15] and grammatical errors [8]. In addition, models have limited input sizes, which makes it difficult to summarize large text corpora. Finally, automatic evaluation metrics for summarization such as the ROUGE score [11] have many shortcomings, penalizing predictions that are different in wording but nevertheless semantically analogous to the reference summary [5].

Many existing work have attempted to address these challenges. For instance, there have been existing approaches to improving abstractive summarization on large text corpora using extractive summarization techniques that can reduce the input size. Notwithstanding, these approaches have yielded considerable success, reaching or even surpassing state-of-the-art performances on their respective data. However, for such models to be used effectively and efficiently in practice, it is important to know how they perform relative to each other and the costs associated with each model. To our knowledge, little to no work has compared the performance of different models using one dataset of focus. Moreover, no previous work has analyzed the computational costs associated with different models.

In this work, we use two different transformer models to study ways to enhance the performance of abstractive summarization on large text corpora. First, we perform experiments to investigate how reducing input sizes impacts the performance of summarization. Then, we evaluate performance using both traditional metrics like the ROUGE score and human evaluation as well as a new learned

evaluation metric BLEURT [18]. Finally, we analyze the differences in time and computational resources across our experiments.

For our experiments, we used the Multi-News [4] text dataset as we found that the data examples were very large. To reduce the input sizes, we preprocessed Multi-News using the Maximum and Word Frequency algorithms to generate compressed datasets. We fine-tune and evaluate the 60-million parameter model T5-small [17] on the different preprocessed datasets. Thereafter, we compare the performance of T5-small against the state-of-the-art 568-million parameter PEGASUS Multi-News [22] model. We conclude our study with a human evaluation of our output predictions followed by a memory and time analysis of PEGASUS and T5.

Our results reveal that fine-tuning T5-small on a Word Frequency preprocessed dataset outperforms PEGASUS in the BLEURT and ROUGE-2 score by 0.045 and 7.5 respectively, when evaluating on a small subset of 100 test examples. These results are also supported by qualitative analysis of the examples. Moreover, we find that T5-small takes substantially less run-time and memory than PEGASUS during the experiments. We conclude that the use of extractive summarization techniques can increase the performance of abstractive summarization on large text corpora.

2. Background

2.1. Existing Problems

Recently, abstractive summarization models have improved with great success, especially in terms of quantitative metrics. However, there are still many existing problems with these models: poor fluency and grammar, factual inconsistencies, limited input sizes, and weak evaluation metrics.

The generation of fluent summaries that are gramatically correct remains a challenge. As Lin et al. (2018) notes, attention-based sequence to sequence models "for abstractive summarization can suffer from repetition and semantic irrelevance" resulting in grammar issues in the output [12]. Table 1 below illustrates an example of a summary that suffers from such issues, which resembles what is often observed in practice.

Gold Reference	Julian went to the store to buy eggs.
Prediction	Julian went to the store, went to the store to buy eggs.

Table 1: An example of a prediction output that suffers from fluency and grammar issues. In particular, the phrase "went to the store" is repeated twice.

Besides poor fluency and grammar, the low degree of factual correctness of a prediction to the reference summary is also a common issue. In Cao et al. (2017), the authors performed a study that analyzed the extent to which abstractive summaries were "faithful" to the original source text and found that 30% of generated summaries by the state-of-the-art abstractive summarization models contained fabricated information [2]. This factual inconsistency is further supported by Nan et al. (2021), which concluded that model predictions suffer from "entity hallucination," a phonemenon where entities or relations between entities absent in the source input appear in the prediction [15]. Similar to previous two studies, Lebanoff et al. (2019) analyzed the outputs of summaries formed by "sentence fusion," where information is pieced together from more than one sentence [8]. From their study, they discovered a large portion of system outputs failed to remain faithful to the original summary [8]. Collectively, these findings point to the need to make systems more faithful, especially as factual correctness is largely regarded as an essential prerequisite to successful real-world summarization systems [2].

Besides the quality of summary predictions, all transformers models have a maximum number of tokens they can take as input, which is often 512 or 1024. In the context of long document summarization, this is problematic for several reasons. First, due the input constraint of these models, most of research in summarization has primarily involved short document datasets as opposed to long documents [6]. Moreover, with summarization of long documents, the input is often truncated, potentially eliminating important context for the model. In Mutasodirin and Prasojo (2021), the authors suggested that using extractive summarization techniques as a shortening strategy can be beneficial in large corpora with salient information throughout the document [14].

Gold Reference: Eating healthy is essential to good health.	
ROUGE-1	**Prediction**
0.857	Eating well is essential to good health.
0.230	In order to be healthy, you need to eat a good, balanced diet.
0.625	Good health is essential to a good career.

Table 2: An example illustrating the shortcomings of the ROUGE score. The second prediction is closer in meaning to the gold reference than the third, but ranks lower in ROUGE-1.

Finally, the automatic evaluation metrics used for abstractive summarization such as the ROUGE score are often not accurate or reliable. The ROUGE score captures the n-gram lexical overlap between the prediction and reference summary [11]. However, due to the diversity of valid ways a source text can be summarized, such metrics which only measure lexical overlaps may unfairly penalize predictions that are different in lexical structure yet quite similar in meaning to the true reference; conversely, the metrics are also prone to unfairly rewarding predictions that are similar in lexical structure but drastically different in meaning. Table 2 provides a concrete example demonstrating the shortcomings of the ROUGE.

Finally, there are large computation and storage costs associated to training and fine-tuning large models. In particular, the training of very large models can take several weeks and the pre-trained models can take hundreds of gigabytes to store [23].

2.2. Abstractive Summarization Models

Currently, many transformer models have been developed for abstractive summarization. In Raffel et al. (2019), the authors explore a variety of transfer learning techniques to create a model called Text-To-Text Transfer Transformer (T5) [17]. Importantly, during pre-training, the authors add a text-specific prefix to the input, which allows the model to learn and distinguish between different text-to-text tasks [17]. The authors used a massive corpora called Colossal Clean Crawled Corpus (C4) to pre-train T5, discovering that element masking yielded strong results [17]. Finally, T5 is

unique in that in uses the same set of parameters to perform a variety of text-to-text tasks [17]. In Zhang et al. (2019), the authors used sentence masking as a pre-training objective to create a summarization transformer-based model called PEGASUS [22]. Unlike previous models like Bidirectional Encoder Representations from Transformers (BERT) [3] which mask words or phrases for pre-training, PEGASUS masks and predicts of important sentences in the source text (or what authors label as "Gap Sentences Generation"), which makes language understanding and generation more effective [22]. PEGASUS is pre-trained separately using the C4 [17] and HugeNews corpus [22]. Because PEGASUS was fine-tuned on many downstream summarization tasks including long document datasets, the authors suggested that increasing the maximum input length could improve performance [22]. Similarly, in Liao et al. (2019), the authors explore ways to improve text summarization on long documents and propose an aggregation mechanism that can be used with transformer models to improve the quality of summaries [10].

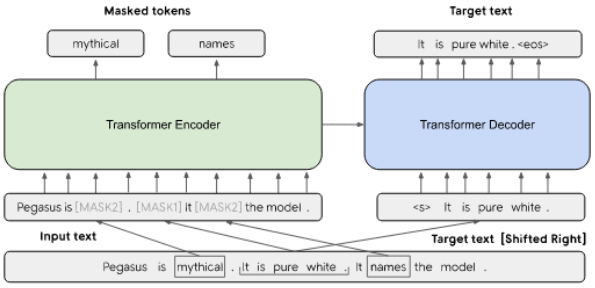

Figure 1: A diagram illustrating the architecture of PEGASUS. The model uses masking at both the token and sentence level as a self-supervised pre-training objective. Image Credit: [22]

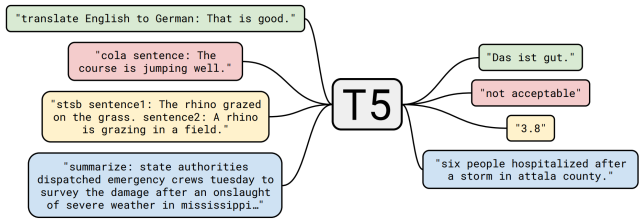

Figure 2: A diagram illustrating the architecture of T5 model. The model takes inputs preprended with a task-specific prefix (e.g., "summarize") and uses the same set of parameters to perform multiple text-to-task tasks. Image Credit: [17]

Collectively, these works demonstrate improvements in summarization. However, as existing works have used different datasets to measure performance, it is difficult to gauge the performance of each model relative to other well-performing models. In addition, all aforementioned works use ROUGE metric as the only automatic evaluation metric, which is subject to the limitations outlined in Section 2.1. To our knowledge, our work will be the first to do a cross-comparative analysis of the performance and behavior of PEGASUS and T5 on the same dataset, using ROUGE, human evaluation, as well as the new learned evaluation metric BLEURT.

2.3. Extractive-Abstractive Approaches

There have been several works that attempt to reduce the input size using a combination of extractive and abstractive techniques. One closely related work to ours is Li et al. (2021), in which the authors propose an extractive-abstractive summarization with explanations (EASE) framework, which masks elements from the source document and then feeds the output for abstractive summarization [9]. Related to EASE, the authors of Tretyak et al. (2020) use pre-trained transformer models for an extractor and abstractor and employ a extractive-abstractive approach for summarization on the long-document scientific journal dataset [20]. Achieving considerable results, both works are related to ours in that they attempt to abstractive and extractive techniques on long documents. However, unlike EASE and the extractor-abstractor which involve the creation of new frameworks, our study *compares* the quantitative and qualitative performance and resource usage of two state-of-the-art models on a single long-document dataset Multi-News [4]. In particular, we investigate whether applying extractive approaches to reducing input size improves the quality of summarization.

3. Approach

We focus on a single multi-document dataset Multi-News [4] in order to study abstractive summarization on large text corpora. We specifically chose Multi-News as opposed to other large-document datasets because the authors of PEGASUS reached a state-of-the-art performance on Multi-News [22] and thus, we use their results as a baseline. Moreover, the authors of PEGASUS suggest that there may be additional ways further enhance performance on this dataset, including a two-stage extractive-abstractive approach [22, 13].

As such, we develop a variety of preprocessing techniques in attempts to reduce the input size. We hypothesize that given the length and multi-document feature of Multi-News, condensing the inputs can extract the most relevant, important information that is needed as context for the model and lead to summaries that are more adequate and meaningful. As automatic evaluation metrics like ROUGE are often used to assess summarization but are not always insightful, we also conjecture that the quality of summaries generated by smaller models need not be worse than the quality of

those generated by larger ones; indeed, they could be comparable or even better.

Given our time and storage constraints, we choose to compare the smaller T5-small to the much larger PEGASUS. In particular, we fine-tune and evaluate T5-small on a variety of datasets we preprocessed. Moreover, we evaluate PEGASUS Multi-News, a PEGASUS model that was fine-tuned on Multi-News. For evaluation, in addition to ROUGE and human evaluation, we employ the recently developed BLEURT score, a more reliable alternative to ROUGE. We conclude our study by measuring the time and memory costs for fine-tuning the two models separately.

To summarize our contributions:

- We use many preprocessing techniques to reduce the input size of the Multi-News dataset
- We conduct numerous experiments using PEGASUS and T5 to generate predictions
- We use the new BLEURT score to evaluate performance between the PEGASUS and T5
- We assess differences in time and computational resources across our models used in our fine-tuning experiments

4. Implementation

4.1. Dataset

For our dataset, we used Multi-News, a collection of news articles gathered from the website newser.com and their corresponding summaries written by professional editors [4]. Multi-News is a multi-document dataset, namely that each example consists of a collection of news article documents that pertain to the same topic. To distinguish between the different articles within a data instance, the articles are delimited by a special token 'lllll'. We show a sample data instance below in Table 3.

Document	A disabled man who spent more than half an hour trapped in Disneyland's "It's a Small World" ride in 2009 has won $8,000 in damages from the amusement park, the man's lawyer said Tuesday. Jose Martinez, a resident of San Pedro (Los Angeles County) who is in early 50s, was stuck in the "Goodbye Room" when the ride broke down the day after Thanksgiving in 2009, said David Geffen, a Santa Ana attorney. Disneyland employees evacuated other riders but had no way to help Martinez, who is paralyzed and uses a wheelchair, Geffen said. Martinez suffers from panic attacks and high blood pressure, both of which became an issue as he sat in the boat, the "Small World" song playing over and over and over, Geffen said. "He was half in the cave of the ride and half out," Geffen said. "The music was blaring. They couldn't get it to go off." Disneyland employees should have called firefighters to evacuate Martinez, but instead they waited for the ride to be fixed, Geffen said. Martinez was eventually treated at a Disneyland first aid station, the lawyer said. Besides failing to take proper care of Martinez while he was stuck on the ride, Disneyland did not notify disabled riders that if "It's a Small World" broke down, they could be trapped, U.S. District Judge James Selna ruled Friday. Martinez sued Disneyland in February 2011 in U.S. District Court in Santa Ana. **...** "The court's saying that this kind of injury is foreseeable and that (Disneyland) has a duty to warn people. Follow @WillKane ||||| (credit: CBS) ANAHEIM (CBSLA.com) — A disabled man was awarded $8,000 from Disneyland in a federal lawsuit after he became stuck for 30 minutes on the "It's A Small World" ride. U.S. District Judge James Selna ruled in favor, in part, for Jose. R. Martinez and his wife, Christina Buchanan-Martinez. According to court records, the judge had also earlier ruled in favor, in part, for Disneyland Resort. The trial focused on disabled access to Disneyland's first aid station **...** All Rights Reserved. This material may not be published, broadcast, rewritten, or redistributed. Wire services contributed to this report.) |||||
Summary	– A disabled California man who spent a hellish half-hour stuck on Disneyland's "It's a Small World" ride as the theme song blared over and over again has been awarded $8,000 in damages. The man, who is in his early 50s and is paralyzed from a spinal cord injury, was left in the "Goodbye Room" when the ride broke down and other riders were evacuated from the boat, the San Francisco Chronicle reports. The manś lawyer says his client suffers from panic attacks and high blood pressure—and a full bladder made the situation even worse. The judge awarded the man $4,000 for "pain and suffering" and another $4,000 for a violation of the Americans with Disabilities Act. "I find a breach of the common-law duty to provide safe premises," the judge said in his ruling, saying Disney has a duty to warn disabled visitors that they could be trapped for an extended period of time when rides break down, CBS reports.

Table 3: A sample example consisting of a document-summary pair from the Multi-News dataset. The document is a multi-document comprised of two articles, which are delimited by the '\|\|\|\|\|' token. The document displayed in the table is truncated as indicated by the ellipsis for illustration purposes.

Multi-News consists of a total of 56216 document-summary pairs. Among them, there are 44972 training, 5622 validation, and 5622 test examples.

We accessed Multi-News through the Hugging Face datasets library. We chose Multi-News because the document length for each data example is very large, making the dataset was suitable for our study on large text corpora. After exploring the dataset, we found that the mean word count of 1792 across all documents was well beyond the maximum input token length of transformer models. Moreover, the word count of the longest document is 449622. Figure 3 below shows the distribution of the lengths of the documents in the dataset along with the mean length. For visualization purposes, we include only documents that have a word count of up to 10000.

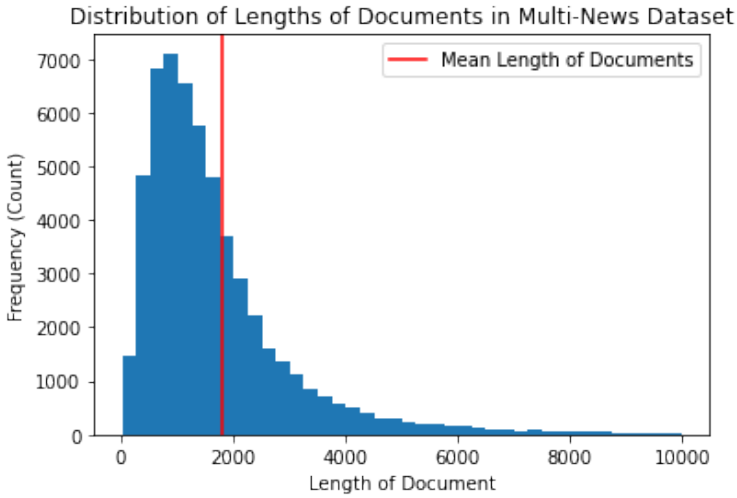

Figure 3: The distribution of lengths of documents in the Multi-News dataset for documents with word count of 10000 or less. The mean length is 1792.

4.2. Model Selection

The two models we used are PEGASUS Multi-News and T5-small.

PEGASUS Multi-News is a encoder-decoder transformer model that was pre-trained on both the C4 and HugeNews corpus using a batch size of 8192 and subsequently fine-tuned on the Multi-News dataset [22]. PEGASUS Multi-News contains 568 million parameters and has a maximum input length of 1024 tokens. In addition, the model contains 16 layers for both the encoder and decoder, 16 self-attention heads, and a hidden size of 1024 [22]. Importantly, PEGASUS Multi-News reached state-of-the-art performances, achieving a best ROUGE-1 and ROUGE-2 score 47.52 and 18.72 [22].

On the other hand, T5-small is a base model that was pre-trained using the C4 corpus but not fine-tuned on a downstream task [17]. T5-small contains 60 million parameters, has a maximum input length of 512 tokens, and contains 6 attention modules [17].

We chose PEGASUS because it is a state-of-the-art transformer model on the Multi-News dataset that we used as a baseline for ROUGE performance. Then, we chose T5-small over other models such as BART since T5-small was a small model that was easy to train and fine-tune.

4.3. Preprocessing Algorithms

As illustrated in Section 4.1, the input size for the dataset is very large. To condense the input size of our dataset, we employed two algorithms, namely the Maximum algorithm and Word Frequency algorithm [16].

4.3.1. Maximum Algorithm

The maximum algorithm takes as input a data example, namely a multi-document, and outputs the longest document in the multi-document. More formally, let $C^{(i)} = \{D_1^{(i)}, ..., D_n^{(i)}\}$ be the i^{th} data instance. Note that $C^{(i)}$ is a multi-document, namely a set of n related documents. In addition, let $L(S)$ be the word count of a text input S. The maximum algorithm $M(C)$ will output the maximum length document in C:

$$M(C) = \underset{D \in C}{\operatorname{argmax}}\{L(D)\} \tag{1}$$

We reasoned that in theory, using the maximum length document could provide a more coherent input. For instance, suppose for some $C^{(i)}$, $D_1^{(i)}$ and $D_2^{(i)}$ are very short and $M(C^{(i)}) = D_3^{(i)}$. In this case, it might be beneficial to use $D_3^{(i)}$ for our model input since it provides additional details that the first two documents doesn't include. As such, we hypothesized that using the Maximum algorithm would improve the "adequacy" of the predictions, namely the extent to which it covers the key points of the multi-document.

4.3.2. Word Frequency Algorithm

We adapted the work of Akash Panchal [16] to implement the Word Frequency (WF) algorithm. The algorithm takes as input a multi-document and outputs an extractive summary, which we use as our new input for abstractive summarization. We implemented two variants: sentence-level WF and paragraph level WF.

First, we define the sentence-level WF. Let $T^{(i)}$ be the i^{th} data instance. Then:

1. If $L(T^{(i)}) \leq 512$, output $T^{(i)}$. Otherwise, proceed to step 2.
2. Split $T^{(i)}$ into m sentences $s_1^{(i)}, ..., s_m^{(i)}$.
3. Create a frequency table F of non stop words in $T^{(i)}$.

4. For each $j \in [1,m]$, compute a score $q(s_j^{(i)})$, the normalized sum of the frequency of non-stop words in $s_j^{(i)}$ according to F. More concretely, $q(s_j^{(i)}) = \frac{F(s_j^{(i)})}{L(s_j^{(i)})}$. Also, let Q be the set of all scores $q(s_j^{(i)})$.
5. Compute the average A of scores in Q. In particular, compute $A = \frac{1}{m}\sum_{j=1}^{m} q(s_j^{(i)})$.
6. Define the threshold $\gamma = 1 + 0.1 \cdot \frac{L(T^{(i)})}{512}$.
7. Initialize O to $s_1^{(i)}$ (that is, always keep the first sentence). Go through all sentences 2 to m. If $q(s_j^{(i)}) > \gamma \cdot A$, concatenate $s_j^{(i)}$ to O. Output O as the summary.

In step 7, we always added the first sentence since empirically, we found that the first sentence provides essential context.

With sentence-level WF, we paired it with Maximum algorithm. Thus, we let $T^{(i)} = M(C^{(i)})$. Importantly, we reasoned that condensing the length of the dataset would reduce the input size and yet still retain the important information. However, we examined the outputs and observed that they were often not coherent and many sentences were out of context.

To increase the coherence of the summary, we also subsequently implemented a paragraph-level WF. We define the paragraph-level WF as follows. Let $C^{(i)}$ be the i^{th} data instance. Then:

1. Split $C^{(i)}$ into k paragraphs $p_1^{(i)}, ..., p_k^{(i)}$, where each paragraph consists of sets of three sentences. Formally, $p_j^{(i)} = \{s_j^{(i)}, s_{j+1}^{(i)}, s_{j+2}^{(i)}\}$.
2. Create a frequency table F of non stop words in $C^{(i)}$.
3. For each $j \in [1,k]$, compute a score $q(p_j^{(i)})$, the sum of the frequency of non-stop words in $p_j^{(i)}$ according to F. More concretely, $q(p_j^{(i)}) = \frac{F(p_j^{(i)})}{L(p_j^{(i)})}$. Also, let Q be the set of all scores $q(p_j^{(i)})$.
4. Compute the average A of scores in Q. In particular, compute $A = \frac{1}{k}\sum_{j=1}^{k} q(p_j^{(i)})$.
5. Initialize the threshold γ to some constant.
6. Initialize O to $s_1^{(i)}$ (that is, always keep the first paragraph). Go through all paragraphs 2 to m. If $q(p_j^{(i)}) > \gamma \cdot A$, concatenate $p_j^{(i)}$ to O. Output O as the summary.

Empirically, we tested various values for γ and found that $\gamma = 0.9$ produced summaries that

balanced both concision and coherence. Thus, we used the dataset preprocessed with $\gamma = 0.9$ for our experiments.

4.4. Pipelines

After preprocessing of the datasets, we fine-tuned T5-small. To prepare the data, we first prepended all the documents in the dataset with the summarization task-specific prefix "summarize" as this was done during the pre-training of the original model. The result was then fed into a T5 Tokenizer using padding and a maximum length of 512. In addition, we set the pad tokens in the label to -100 for proper computation of the loss during training and evaluation. The T5-small model, specifically initialized for conditional generation, takes in the tokenized input and outputs a tokenized output. Finally, we used the tokenizer to decode the tokenized output into a text summary. During this decoding step, we applied a repetition penalty to prevent repetitions in the summary. After fine-tuning, we then evaluated the predictions using a small random sample of 100 examples from the test set. During training, we used a batch size of 4, the optimizer Adafactor, and a learning rate of 10^{-4}. We illustrate the T5-small pipeline in Figure 4 below.

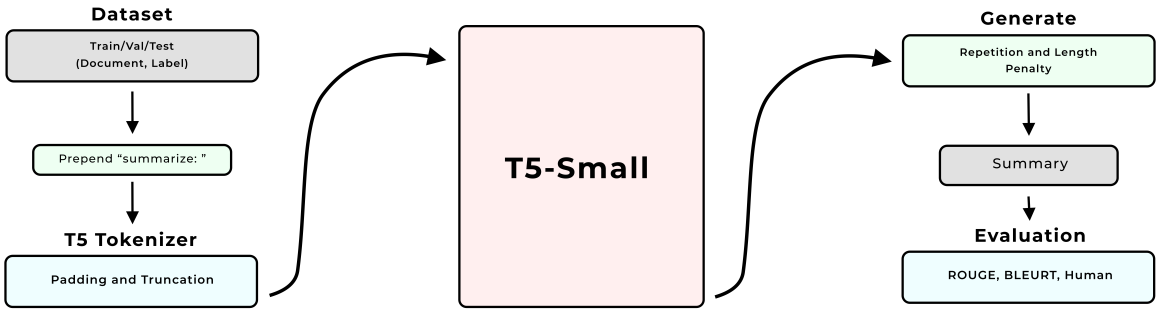

Figure 4: The pipeline for the fine-tuning and evaluation of the T5-small model. After preprocessing and tokenization, the T5-small model is trained and the decoded generated summaries undergo evaluation.

Besides fine-tuning T5-small, we also evaluated PEGASUS Multi-News on the same examples in the sample of test data used in the T5-small experiments. Like with T5-small, we used a tokenizer to tokenize and decode the input. In this case, we used the PEGASUS Tokenizer. We illustrate the PEGASUS pipeline in Figure 5 below.

Figure 5: The pipeline for the evaluation of the PEGASUS Multi-News model. The generated summaries undergo evaluation.

For training and evaluation, we use one NVIDIA Tesla K80 GPU. Our experiments were performed using Google Colab, Microsoft Azure, and the Hugging Face library.

4.5. Experiments

In total, we performed five separate experiments that test a different combination of the model and preprocessing type. For fine-tuning, we used all the training and validation data in the Multi-News dataset. For evaluation, we selected a random sample of 100 examples from the test dataset. To define our experiments, we denote each experiment by an acronym:

1. **T5-MWF**: T5-MWF stands for "T5-small on the **M**aximum **W**ord **F**requency dataset". In particular, we preprocessed Multi-News using the Maximum Algorithm to get a dataset A. Then, we apply the sentence-level Word Frequency algorithm for each maximum length document in A to get B. Using the resulting shortened dataset B, we fine-tuned T5-small for 2 epochs.
2. **T5-MOD**: T5-MOD stands for "T5-small on the **M**aximum of the **O**riginal **D**ocuments". In particular, we preprocessed Multi-News using the Maximum Algorithm to get a dataset C. Using the resulting shortened dataset C, we fine-tuned T5-small for 2 epochs.
3. **T5-VOD**: T5-VOD stands for "T5-small on the **V**anilla **O**riginal **D**ocuments". Vanilla Original Documents means that we used the original dataset D and no preprocessing was performed. Using D, we fine-tuned T5-small for 2 epochs.
4. **T5-VWF**: T5-VOD stands for "T5-small on the **V**anilla **W**ord **F**requency dataset". In particular, we applied the paragraph-level Word Frequency algorithm on the original "vanilla" dataset to get a resulting dataset E. Using E, we fine-tuned T5-small for 2 epochs.

5. **PEGASUS-VOD**: PEGASUS-VOD stands for "PEGASUS Multi-News on the **V**anilla **O**riginal **D**ocuments dataset". In particular, we simply used PEGASUS Multi-News to evaluate the original "vanilla" dataset. No fine-tuning was performed.

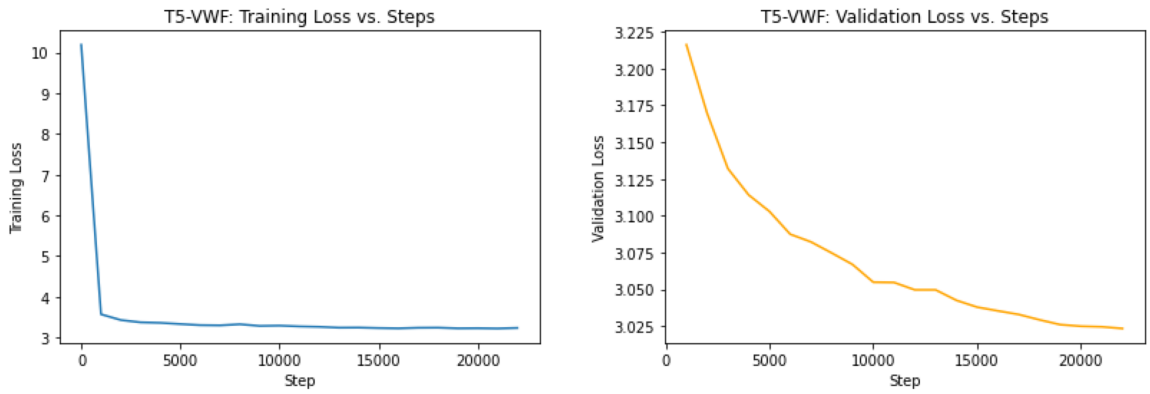

Figure 6: Training (left) and validation (right) loss curves for the T5-VWF experiment. In particular, training loss converges after a few thousands steps whereas the validation loss starts to converge after about 20000 steps.

Figure 6 above depicts the training and validation curve for the T5-VWF experiment. Furthermore, we summarize details of the experiments in Table 4 below.

Name	Model	Preprocessing	Fine-tuning	Number of Epochs
T5-MWF	T5-small	Maximum and Word Frequency (Sentence-Level)	Yes	2
T5-MOD	T5-small	Maximum	Yes	2
T5-VOD	T5-small	None	Yes	2
T5-VWF	T5-small	Maximum and Word Frequency (Paragraph Level)	Yes	2
PEGASUS-VOD	PEGASUS Multi-News	None	No	–

Table 4: The five experiments performed. All T5-small experiments are fine-tuned for 2 epochs.

5. Results

5.1. Evaluation Metrics

We employ three evaluation metrics: the ROUGE metric, the BLEURT score, and human evaluation.

5.1.1. ROUGE

The ROUGE metric computes the n-gram overlap between the reference and predicted summaries [11]. At test time, we randomly selected 100 test documents and labels. Using the documents, we generated predicted summaries and computed the ROUGE-1 and ROUGE-2 score. The ROUGE score has a range between 0 and 100, with a higher number being indicative of a stronger prediction.

5.1.2. BLEURT

As noted in Section 2.1, the ROUGE metric can be limited in that it chooses not to reward a summaries that are semantically accurate but perhaps not lexically similar to the reference. To better address this issue, we also chose to use the novel evaluation metric BLEURT [18]. Unlike ROUGE, the BLEURT score is a *learned* evaluation metric. To help the model learn the different ways to express the same idea in a summary, the BLEURT metric is pretrained using BERT on millions of synthetic examples generated by "randomly perturbing" text from Wikipedia [18]. Following pretraining, the BERT model is subsequently fine-tuned on the news-domain WMT Metrics Shared Task dataset with over 260,000 human ratings [19, 18]. According to the authors of BLEURT, the pretraining is essential as it makes the metric more robust to changes in the domain or quality of the dataset used for evaluation [19, 18]. Importantly, as the BLEURT score is trained on human evaluations, the metric has a stronger correlation with human judgments [19]. The BLEURT score has a range between -1 and 1, with a higher number being indicative of a stronger prediction.

At test time, we used the same 100 random test documents and labels used to compute the BLEURT scores. We then averaged the scores together to compute the final BLEURT score.

5.1.3. Human Evaluation

We also manually evaluated the generated predictions using the set of four criteria as follows.

1. **Fluency and Grammar**: The extent to which the predicted summary is coherent, fluent, and grammatically correct.
2. **Factual Correctness**: The extent to which details in the predicted summary are consistent and true with respect to the reference.
3. **Adequacy**: The extent to which the predicted summary covers the main ideas in the reference.

4. **Relevance**: The extent to which details in the predicted summary are essential and pertinent with respect to the reference.

For each criteria, a score was given from 1 to 5, with 5 being the best. At the very end, we averaged the scores together to compute a final score associated with the prediction.

At test time, we chose 20 random test examples and scored them (by ourselves) according to the four aforementioned criteria. In particular, we first studied the reference summary to extract the key ideas. To systematize this process, we wrote down the who, what, where, when, and why (denote this as the "Five W's") of the summary. Then, we examined the reference for any fluency, grammar, or factual inconsistency issues. We then judged the adequacy of the prediction based on how well it covered the Five W's. Finally, for relevance, we assessed whether the summary contains irrelevant or unwarranted information.

Experiment	BLEURT	ROUGE-1	ROUGE-2	Human Evaluation
T5-MWF	0.3813	34.62	10.51	3.4438
T5-MOD	0.4062	28.54	19.06	3.6000
T5-VOD	0.4146	38.54	14.13	4.0438
T5-VWF	**0.4796**	32.76	**27.50**	3.9813
PEGASUS-VOD	0.4333	**47.49**	20.13	**4.4000**

Table 5: Quantitative results across the five experiments for BLEURT, ROUGE-1, ROUGE-2 and human evaluation scores. The names of the experiments are defined in Section 4.5. For all metrics, a higher number indicates a better performance. The highest numbers for each metric are bolded.

5.2. Quantitative Results

We present quantitative results in Table 5. The results for the ROUGE-1 and ROUGE-2 metric for PEGASUS-VOD (namely 47.49 and 20.13 respectively) are consistent with the original numbers in Zhang et al. (2020), where PEGASUS Multi-News achieved a ROUGE-1 and ROUGE-2 score of 47.52 and 18.72, respectively [22]. Notably, our ROUGE-2 was higher than the original ROUGE-2 by 1.41. This discrepancy is likely due to the fact that the original paper evaluated using the whole test set whereas we used a small sample of test examples for evaluation.

Moreover, surprisingly, our findings also reveal that the T5-VWF model obtained a higher performance than the state-of-the-art PEGASUS in terms of both the BLEURT and ROUGE-2 scores. In particular, the BLEURT for T5-VWF was 0.0463 higher than PEGASUS-VOD, reflecting an over 36% increase. Furthermore, the ROUGE-2 for T5-VWF was 7.37 higher than PEGASUS-VOD, reflecting an over 10% increase. We conjecture that the high performance of T5-VWF is a result of preprocessing the data using Word Frequency at the paragraph level. In particular, extracting and joining together the most important paragraphs seemed to shorten the input without severely compromising the fluency and coherence of the text. We presume that using sets of three sentences preserved the context of the key ideas, which resulted in the shortened dataset being a collection of separate, interdependent paragraphs.

The T5 experiments that involve the Maximum Algorithm exhibit consistently worse performance across all metrics compared to those with the "vanilla" documents. We hypothesize that the low performance of the T5-MWF experiment is a result of preprocessing the data using Word Frequency at the sentence-level. The issue with extractive summarization at the sentence-level is that during the algorithm, a sentence is often removed from its original context and joined together with distant sentences that were not directly related to it; this leads to the condensed text being incoherent and not representative of the original. Indeed, we inspected some of the examples in the T5-MWF dataset and observed that many of them were incoherent and incomprehensible. For instance, it was often unclear how a given sentence was related to its context. Likely as a result of such issues, the T5-MWF generated summaries that often were not fluent nor adequate in meaning, which is further analyzed in Section 5.3.

Moreover, the results for T5-MOD was also contrary to our original expectation that applying the Maximum algorithm may improve the adequacy of the prediction. In reality, however, it was likely that as the maximum input was still very long, the input was not "condensed" enough to successfully insert all the essential details into the model. As such, applying the Maximum Algorithm likely *decreased* the adequacy of the model relative to just using the original document as input. This is evidenced by T5-VOD outperforming T5-MOD in all metrics with the exception of ROUGE-2.

Indeed, among the 20 test examples we use for human evaluation, the average adequacy score for T5-VOD of 3.775 was higher than the adequacy score for T5-MOD of 3.4. This suggests that the beginning of the multi-document may already contain most of the essential details needed for a good quality summary. In the end, we conclude that condensing the dataset using the Maximum algorithm is a naive approach for text shortening and is not conducive to enhancing the performance.

Finally, PEGASUS-VOD achieves the highest ROUGE-1 and Human Evaluation scores. We attribute the higher human evaluation performance of PEGASUS-VOD to the fact that PEGASUS is a much larger model with a larger input size, which results in better summaries overall based on the four set of criteria defined in Section 5.1.

Ground Truth	What's a newly divorced Russian strongman to do when most of the world wants to throttle him and he's getting no respect from his own supporters? Play a little hockey, that's what. And pretty badly, reports the Week in a look at Vladimir Putin's foray onto the ice yesterday at something called the "All-Russia Festival of Amateur Teams of the Night Hockey League" in Sochi. Even with a less-than-motivated opposition goalie and near-assists from NHL players, Putin choked several times. Yet somehow, reports the Daily Mail via the Voice of Russia, Putin managed to score six times and had five assists. Despite his, er, heroics, Putin's team eventually lost, but no hard feelings, says the president: "There are no winners or losers here. This is a friendly game." Click for the Week's GIF of Putin attempting to put his helmet on—backward.
T5-MWF	A team-mate who shot the puck into the net was apparently eager to help him, reports the BBC. Putin's side were knocked out by Finland at the quarter-final stage of the tournament, but they aren't sure how long it will take to get back on track. "It's not clear what happened," said one player in his final game. "We don't know why." The former premier is expected to face Russia in the semi-finals next week.
T5-MOD	The Russian president took to the ice at the All-Russia Festival of Night Hockey League in Sochi today wearing number 11 jersey. He missed several goals, and sporting celebrities nudged in the puck for him instead, until he finally scored Putin often does macho publicity stunts and has shot a whale, hugged a polar bear, and 'found' Greek vases while scuba-diving. Now Vladimir Putin is at it again— playing ice hockey with sporting legends in front of thousands of apparently adoring fans. "It doesn't look like there was much opposition," said the star.
T5-VOD	Vladimir Putin has been playing ice hockey with sporting legends in front of thousands of apparently adoring fans. But the Russian premier, 61, was not without a little help—he missed several goals by a whisker, leaving his celebrity teammates to give the puck a gentle nudge in the right direction, before finally hitting home on his own. The NHL star took a break from crisis talks with the West about his annexing of Crimea to play alongside retired Soviet hockey legends, including Vyacheslav Fetisov and Pavel Bure. "It doesn't look like there was much opposition," said one player.'
T5-VWF	The Russian president took to the ice at the All-Russia Festival of Night Hockey League in Sochi wearing number 11 shirt. Vladimir Putin missed several goals by a whisker, leaving his celebrity teammates to give the puck a gentle nudge in the right direction, before finally hitting home on his own. In at least two attempts, Putin just missed the goal, but the puck went in after a gentle tap from another player. But the upbeat president reportedly told the crowd: "There are no winners or losers here. This is a friendly game. It's a show and everyone enjoyed it. All this [the Night Hockey League] was invented to give an additional boost to the development of sports and physical culture.
PEGASUS-VOD	Vladimir Putin took a break from crisis talks with the West to play some hockey in Sochi—and he scored (.....) His opponents, reportedly amateurs, were reportedly amateurs. In at least two attempts, Putin just missed the goal, but the puck went in after a gentle tap from another player. And in one sequence, shown on the state-linked English-language channel RT, the premier eventually scored, to the delight of the crowd.

Table 6: A comparison of predictions to the ground truth for a sample test example. For PEGASUS-VOD, a portion of text is omitted for display purposes. Blue indicates an issue with the summary.

5.3. Qualitative Results

We provide an example in Table 6 depicting a comparison of predictions to the ground truth. During human evaluation, we first examined the ground truth and described the key points in the form of the Five W's. In the example, we define the Five W's as follows:

1. Who: The Russian President Vladmir Putin
2. What: President Putin took a break from politics and played some hockey
3. Where: Sochi, Russia
4. When: Not specified
5. Why: The article is somewhat satirical in mentioning that Putin needed help to score and that he "choked" several times

For T5-MWF, the prediction does not clearly establish Putin as the individual of focus. In addition, the "him" in the first sentence reflects semantic ambiguity. Overall, the summary lacks fluency and relevance and we gave it an average score of 3.5 out of 5. For T5-MOD, the summary was more relevant and adequate but still suffered issues in fluency and grammar. The "wearing number 11 jersey" misses an indefinite article, details in the middle highlighted sentence is somewhat unwarranted, and it is not clear who "star" is in the last sentence. T5-MOD received an average score of 3.75. T5-VOD generated a fluent, relevant summary covering most of the Five W's. However, the summary contains factually incorrect information as Putin is not an "NHL star". Overall, we gave T5-VOD an average score of 4.5. Similarly, T5-VWF generated a fluent, relevant summary covering nearly all the key points. There were a few issues including the phrase "wearing number 11 shirt" missing an indefinite article and the second and third sentence containing repeated information. We gave T5-VWF an average score of 4.375. Finally, PEGASUS-VOD generated a very adequate and relevant summary yet suffered from some grammar issues. In particular, "reportedly amateurs" is repeated and the last sentence exhibited awkward syntax. We gave PEGASUS-VOD an average score of 4.5.

After human evaluation using such a process, we analyzed the predictions and made several interesting high-level observations. First, we find that PEGASUS-VOD has a tendency to generate

long summaries, which often need to be truncated. In addition, compared to the other models like T5-VOD and T5-VWF, PEGASUS-VOD generates summaries that contain a higher proportion of quotations taken directly from the source document. We hypothesize that this discrepancy may be due to the unique pre-training objective of PEGASUS where the model predicts whole "gap" sentences, which makes it more prone to include quotations. In addition, we find that despite high quantitative performance, a large number of T5-VWF summaries still lack fluency, which could be the result of insufficient context. Lastly, we observe that the generations T5-MWF and T5-MOD were often inadequate and incoherent, which is consistent with our quantitative results.

For reference, we include several additional examples that highlight these observations in the 24.

5.4. Computational Cost Analysis

We chose to compare the computational costs of our best model T5-VWF against PEGASUS. As our resources were limited, we selected only the first 100 training and validation examples. Then, we fine-tuned PEGASUS Large on the "vanilla" dataset and T5-small on the VWF preprocessed dataset for 1 epoch each using one 12GB NVIDIA Tesla K80 GPU. Our findings are as displayed below in Table 7.

Model	Time (sec)	Peak Memory (MiB)
PEGASUS	1044.45	8733.53
T5	**31.97**	**6759.12**

Table 7: Time and memory usage for fine-tuning PEGASUS and T5 for 1 epoch using 100 training and validation examples. Lower numbers are bolded.

As the table depicts, the time and memory it takes to fine-tune T5-small is less than that of PEGASUS. From this, we conclude that the T5 model can generate summaries of quality comparable to their counterparts generated by PEGASUS and can do so using significantly less time and considerably less memory. This is important as it suggests that for summarization systems that have limited computational resources, the use of T5-small could be a reasonable alternative to PEGASUS or other larger transformer models.

6. Conclusion

In this work, we compare the performance of a smaller T5-small model against the larger state-of-the-art PEGASUS model using three evaluation metrics. While some text-shortening techniques did not enhance performance, using the Word Frequency algorithm on a paragraph level did. In particular, following preprocessing, the T5-VWF model surpassed PEGASUS by 7.5 and 0.045 in the ROUGE-2 and BLEURT score respectively. In addition, our findings reveal that T5 model is more computationally efficient than PEGASUS.

6.1. Limitations and Future Work

Our work has many limitations. First, the dataset we chose was a multi-document dataset and it was not clear whether how the documents in each instance were ordered. Thus, our results may have been influenced by potential positional bias that exists in the dataset. Another limitation was that human evaluation may have bias, as we were the sole evaluators.

In the future, we wish to investigate how different adjustments in the preprocessing steps or model architecture can lead to improvements in fluency, relevance, and adequacy. In addition, we would like to include more human evaluators to assess qualitative performance, perhaps using services like Amazon Mechanical Turk. Finally, we believe that our work could be extended to other datasets that are considered large text corpora.

6.2. Acknowledgements

I would like to sincerely thank my adviser Danqi Chen for her guidance and mentorship throughout of my project, the teaching assistants and my peers in the COS IW04 “Hands-on Deep Learning for Language Understanding ” Seminar for their feedback, and the Princeton COS Independent Work staff for their support.

References

[1] S. Babar, M. Tech-Cse, and Rit, "Text summarization:an overview," 10 2013.

[2] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact aware neural abstractive summarization," *CoRR*, vol. abs/1711.04434, 2017. [Online]. Available: http://arxiv.org/abs/1711.04434

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[4] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model," 2019.

[5] A. Gaskell, "On the summarization and evaluation of long documents," Ph.D. dissertation, 2020.

[6] L. Gonzalez, S. Lu, and W. Buchanan, 2021. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/reports/final_reports/report019.pdf

[7] C. Khatri, G. Singh, and N. Parikh, "Abstractive and extractive text summarization using document context vector and recurrent neural networks," *CoRR*, vol. abs/1807.08000, 2018. [Online]. Available: http://arxiv.org/abs/1807.08000

[8] L. Lebanoff, J. Muchovej, F. Dernoncourt, D. S. Kim, S. Kim, W. Chang, and F. Liu, "Analyzing sentence fusion in abstractive summarization," *CoRR*, vol. abs/1910.00203, 2019. [Online]. Available: http://arxiv.org/abs/1910.00203

[9] H. Li, A. Einolghozati, S. Iyer, B. Paranjape, Y. Mehdad, S. Gupta, and M. Ghazvininejad, "EASE: extractive-abstractive summarization with explanations," *CoRR*, vol. abs/2105.06982, 2021. [Online]. Available: https://arxiv.org/abs/2105.06982

[10] P. Liao, C. Zhang, X. Chen, and X. Zhou, "Improving abstractive text summarization with history aggregation," *CoRR*, vol. abs/1912.11046, 2019. [Online]. Available: http://arxiv.org/abs/1912.11046

[11] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[12] J. Lin, X. Sun, S. Ma, and Q. Su, "Global encoding for abstractive summarization," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 163–169. [Online]. Available: https://aclanthology.org/P18-2027

[13] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," *CoRR*, vol. abs/1801.10198, 2018. [Online]. Available: http://arxiv.org/abs/1801.10198

[14] M. A. Mutasodirin and R. E. Prasojo, "Investigating text shortening strategy in bert: Truncation vs summarization," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2021, pp. 1–5.

[15] F. Nan, R. Nallapati, Z. Wang, C. N. dos Santos, H. Zhu, D. Zhang, K. R. McKeown, and B. Xiang, "Entity-level factual consistency of abstractive text summarization," *CoRR*, vol. abs/2102.09130, 2021. [Online]. Available: https://arxiv.org/abs/2102.09130

[16] A. Panchal, "Text summarization in 5 steps using nltk," Feb 2021. [Online]. Available: https://becominghuman.ai/text-summarization-in-5-steps-using-nltk-65b21e352b65

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *CoRR*, vol. abs/1910.10683, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[18] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: learning robust metrics for text generation," *CoRR*, vol. abs/2004.04696, 2020. [Online]. Available: https://arxiv.org/abs/2004.04696

[19] T. Sellam and A. P. Parikh, "Evaluating natural language generation with bleurt," May 2020. [Online]. Available: https://ai.googleblog.com/2020/05/evaluating-natural-language-generation.html

[20] V. Tretyak and D. Stepanov, "Combination of abstractive and extractive approaches for summarization of long scientific texts," *CoRR*, vol. abs/2006.05354, 2020. [Online]. Available: https://arxiv.org/abs/2006.05354

[21] M. Yang, C. Li, Y. Shen, Q. Wu, Z. Zhao, and X. Chen, "Hierarchical human-like deep neural networks for abstractive text summarization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2744–2757, 2021.

[22] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization," *CoRR*, vol. abs/1912.08777, 2019. [Online]. Available: http://arxiv.org/abs/1912.08777

[23] Z. Zhang, Y. Gu, X. Han, S. Chen, C. Xiao, Z. Sun, Y. Yao, F. Qi, J. Guan, P. Ke, Y. Cai, G. Zeng, Z. Tan, Z. Liu, M. Huang, W. Han, Y. Liu, X. Zhu, and M. Sun, "CPM-2: large-scale cost-effective pre-trained language models," *CoRR*, vol. abs/2106.10715, 2021. [Online]. Available: https://arxiv.org/abs/2106.10715

Appendix

Ground Truth	Over the past decade, "Isis" has "become synonymous with 'a better birth experience,'" Georgia's Isis Women's Health Care writes on its website. That's recently changed, for obvious reasons. "One morning we woke up to news of a beheading by a terrorist organization the media chose to refer to as ISIS," the business states. "That meant bad news for us." Since then, Isis Women's Health Care has received dozens of death threats, WSB-TV reports. "I became concerned for me and my staff," Dr. Hughan Frederick tells WXIA. Finally, one caller made it clear it was time for a name change. "This particular individual identified himself as ex-military and indicated for us not to be surprised if someone were to come and shoot up the building," Isis administrator Randy Haviland says. WSB-TV reports Isis Women's Health Care will announce its new name at the end of the month. But judging by its website, it will now be known as Nile Women's Health Care. That's in keeping with its original Egyptian theme—the business was named for the Egyptian goddess of motherhood. Even before a new name was chosen, Isis removed signage from outside its three Atlanta-area offices to head off further threats. "This year marks the 10th anniversary for Isis Women's Health Care, but instead of hoisting celebration banners, we are taking down the namesake that has proudly delivered thousands of babies," the business states. (It's not the only business named Isis in America.)
T5-VOD	A local women's clinic is being forced to change names because people have confused it with a terrorist group, reports the Atlanta Morning Herald. The name ISIS Women's Health Care has received dozens of death threats from callers who thought the clinic was associated with the group that is known for torture, suicide bombings, and executions. "Thankfully, we're not closing our doors, but we are opening a new chapter," says Dr. Hughan Frederick. "We've spent a decade and thousands of dollars perfecting and building our brand in this community. Now it's all out the window. But instead of being sad, we want to turn this into something positive and fun. We've had to undergo a complete image change from pamphlets, to websites, to business cards."
T5-VWF	A woman's clinic in Atlanta has been forced to change names because people have confused it with a terrorist group, reports CNN. The name ISIS Women's Health Care says it received dozens of death threats from callers who thought the clinic was associated with the group that is known for torture, suicide bombings, and executions. "This year marks the 10th anniversary for Isis Women's Health Care, but instead of hoisting celebration banners, we're taking down the namesake that has proudly delivered thousands of babies," Dr. Hughan Frederick tells Fox News. "We're not closing our doors, but we are opening a new chapter." He also chose the name ISIS as its tie to the Egyptian goddess of motherhood. "I'm not going to lie, this is hard."
PEGASUS-VOD	A women's clinic in Atlanta says it's had to change its name because people think it́s associated with ISIS, the Atlanta Journal-Constitution reports. Isis Women's Health Care says it has received death threats and had to undergo an image change from pamphlets, to websites, to business cards. "We've spent a decade and thousands of dollars perfecting and building our brand in this community. Now it's all out the window," says Dr. Hughan Frederick, who picked the name for its tie to the Egyptian goddess of motherhood. "But instead of being sad, we want to turn this into something positive and fun. We've had to undergo a complete image change from pamphlets, to websites, to business cards." Frederick says the name Isis has become synonymous with "a better birth experience" in the Atlanta area. "Through the hard work of so many people the name, Isis, has become synonymous with 'a better birth experience' in the north Metro Atlanta area," he writes on the clinic's Facebook page. "It's a name that we have all become proud to be associated with." Frederick says he and his brother picked the name for the clinic 10 years ago because they wanted to build the best practice they could

Table 8: A comparison of a reference summary and generated summaries. The PEGASUS generates a longer summary than the T5 "Vanilla" models and was truncated.

Ground Truth	A Baltimore woman who reportedly saw a young man she knew (some speculate it's her son) on TV throwing rocks at police during city protests took matters into her own hands—literally, per WMAR. A video shows her yesterday dragging a teen sporting a face-covering sweatshirt away from the crowds, hitting him in the face, and yelling, "Take that f—ing mask off!" and "You wanna be out here doing this stuff?!" The kid skulks away with her hot on his tail, and the video—shown from various angles at Independent Journal Review—went viral on social media, with some, like Fox News contributor Charles Payne, calling her the "mom of the year," per WGN. "My mother would have done the same thing," Payne tweeted. Meanwhile, at a press conference last night, Baltimore Police Commissioner Anthony Batts said, per the Washington Post, "She started smacking him on the head because she was so embarrassed. I wish I had more parents who took charge of their kids tonight." (The Orioles COO has his own controversial take on the protests.)
T5-VOD	A Baltimore mom saw her 16-year-old son throwing rocks at police on television, and the mother repeatedly struck him as he tries to walk away. The video shows the mother repeatedly striking her son, chasing him as he tries to walk away. "That's my only son, and at the end of the day, I don't want him to be a Freddie Gray," she tells CBS News. "But to stand up there and vandalize police officers, that's not justice."
T5-VWF	A Baltimore mom saw her son throwing rocks at police on television, but the video was so widespread that Baltimore Police Commissioner Anthony Batts brought it up in an overnight news conference Monday. "And if you saw in one scene, you had a mother who grabbed their child who had a hood on his head and she started smacking him on the head because she was so embarrassed," Batts told reporters. "I wish I had more parents who took charge of their kids tonight. I think these were youth coming out of the high school and they thought it was cute to throw cinder blocks at the police department and address it that way."
PEGASUS-VOD	Toya Graham says she was worried about her 16-year-old son's safety when she confronted him on the streets of Baltimore during yesterday's unrest. "That's my only son, and at the end of the day, I don't want him to be a Freddie Gray," the single mother of six tells CBS News. "But to stand up there and vandalize police officers, that's not justice." In video of the encounter that went viral, Graham can be seen repeatedly hitting the black-clad teen as he tries to walk away, the Washington Post reports. "That's my only son, and at the end of the day, I don't want him to be a Freddie Gray," Graham tells CBS. "But to stand up there and vandalize police officers, that's not justice." Baltimore Police Commissioner Anthony Batts says he wishes more parents "took charge" of their kids during the unrest. "If you saw in one scene, you had a mother who grabbed their child who had a hood on his head and she started smacking him on the head because she was so embarrassed," he says. "I wish I had more parents who took charge of their kids tonight. I think these were youth coming

Table 9: A sample example in which we gave the summary from T5-VWF a higher score than the summary from PEGASUS. In particular, PEGASUS exhibited repetition and incoherence.