Multi-document Summarization with 2-Stage Transformers

University of California, Berkeley | Summer 2021 W266 Final Project

Duncan Howard Julian Tsang

Abstract

We perform multi-document summarization using a two-stage transformer pipeline constructed of extractive and abstractive steps on the Multi-News dataset. Transformers with standard self-attention mechanisms are limited in input document length due to quadratic memory and computation scaling, forcing text to be truncated and preventing the summarization of long documents and multiple input documents. We use a two-stage approach that takes advantage of full self-attention mechanism models by summarizing individual input documents before passing to an abstractive model for a complete summary. We show and compare these results against our baselines, as well as another model that does not perform truncation and incorporates a modified attention mechanism.

1. Introduction

Multi-document summarization provides a range of real-world applications, such as generating convenient summaries on news, reviews, emails, legal documents, etc. As media becomes even more readily available online, the volume that readers face can become overwhelming. Thus, summarization can be used as a tool to reduce information overload and retain the most critical information to present to a reader. Multi-document summarization offers additional utility by synthesizing information from different sources and in different formats.

Summarizing multiple documents presents challenges beyond ordinary summarization tasks because the volume of input documents can become vast and the relationship between them complex. Documents written about the same topic could provide complementary, overlapping, and even conflicting information. Furthermore, it is challenging for models to identify the most important pieces of information, while simultaneously generating clear, concise, non-redundant, factual, and readable summaries (Ma et al. 2020).

Although transformers have shown significant ability to produce cohesive text summaries of individual documents, longer sequences are challenging to process because memory and computation requirements scale quadratically with input length in models that utilize standard "full" self-attention mechanisms.

In this work, to circumvent the limitations associated with full self-attention, we will explore various methods of utilizing pre-trained Transformer based models for the multi-document summarization task, including implementing new ensemble models, where we take the outputs of one model and use them as inputs into a second model. The intuition behind this follows the human approach to summarizing multiple documents. For example, when faced with three documents to perform summarization, a person would likely read each of the three documents, find the lines most relevant to the topic by extracting facts and quotes from each document, and then reword them together to assemble a final abstractive output. We compare our results with the Longformer-Encoder-Decoder model(Beltagy, Peters, and Cohan 2020), which uses a modified self-attention mechanism and does not require significant text truncation.

2. Related Work

Summarization is a sequence-to-sequence task, taking input text to return another set of output text. Encoder-decoder based models are a natural fit for this task as encoders embed documents into hidden

representations containing semantic and/or syntactic information. Then, the decoder processes the latent embeddings to generate summaries. A lot of work has been done using pre-trained language model encoders like BERT and RoBERTa (Liu et al. 2019), which have a lot of contextual knowledge of the English language. For example, BERT has been used to build an extractive summarizer (Miller 2019) in conjunction with K-Means clustering to find the most relevant sentences to the topic cluster centroids.

More recent work has utilized transformer language models in the summarization task, especially in the multi-document setting. BART (Lewis et al. 2019) has shown to be particularly effective in text generation and in comprehension tasks, using few phrases directly from the source text and inferring evidence to produce abstractive summaries. Hybrid summarization has been explored as well, combining extractive and abstractive approaches. Di Fabbrizio, Stent, and Gaizauskas (2014) utilized an extractive-abstractive approach for single-document summarization and argued that the hybrid approach allowed for more readable and compact summaries. Subramanian et al. (2019) utilized a LSTM model for their extractive model and coupled it with a transformer trained to output abstractive summaries. Liu et al. (2018) developed a model that selects the top-K tokens to train a transformer based decoder-only abstractive model that produced multi-paragraph Wikipedia-like summaries. In this approach, selecting the top-K tokens is akin to extractive summarization. Taking a step further in the hybrid summarization setting, Hokamp et al. (2020) implements a pre-trained encoder-decoder transformer and treats multi-document summarization as a single-document problem. They follow a MapReduce-like structure by mapping inputs of relevant documents into the same model and reducing summary outputs into a single summary. We adopt a similar approach and ensemble two pre-trained transformer models, where we generate summaries from a first model and combine them before feeding them into a second model.

Several papers have been explored to reduce the memory and computation costs required by standard transformer models, including two-stage processing, and truncating or chunking input text. Recent models, including BigBird (Zaheer et al. 2020) and Longformer (Beltagy, Peters, and Cohan 2020), have sparsified the self-attention mechanism to reduce time complexity while preserving full-document attention.

3. Dataset

We have identified the Multi-News Dataset (Fabbri et al. 2019), a large-scale multi-document summarization news dataset that contains 56,216 articles-summary pairs. This is contrary to datasets referenced in past papers in this topic, which were largely performed on smaller datasets with less than 100 articles-summary pairs like DUC 2004[1] and TAC 2011[2]. The larger dataset that Multi-News provides will allow for more sophisticated deep learning methods. The dataset is made available via Google Drive download and via Huggingface.

The Multi-News dataset consists of news articles from newser.com along with professionally written summaries. Within the 56,216 topics covered within the dataset, there are 156,266 total articles. We cleaned the dataset by removing articles generated by errors in the Multi-News webcrawler. Some article-summary sets consisted entirely of faulty articles and were removed completely, while other sets had a mix of faulty and standard articles. Summaries with either one or no viable articles remaining were removed as well. Our final dataset results in 51,421 topics with 136,766 total articles. Within each topic, articles are concatenated together with a <story_separator_special_tag>. The average article length is 706 words and the average summary length is 220 words. Refer to Appendix C for examples of faulty articles.

4. Approach

The Transformer architecture leverages the self-attention mechanism and retains long-range dependencies. Still, transformers have context windows which can limit their ability to summarize over longer documents. Using a combination of extractive and/or abstractive methods is our attempt to address this problem. Extractive and abstractive models would be trained separately, and then be sequentially ensembled to perform

[1] https://duc.nist.gov/
[2] https://tac.nist.gov/

summarization. We illustrate the following combinations of ensemble approaches: extractive-abstractive and abstractive-abstractive. For example, in our extractive-abstractive approach, sentences are first fed into a BERT-based extractive summarizer to generate the most important sentences. These sentences are subsequently fed into an abstractive transformer model, which then outputs an abstractive summary. Performing the extractive step allows the abstractive model to focus on the important sentences in the document, accommodating for the limited context window.

Models

BERT/RoBERTa. (Liu et al. 2019) For our abstractive baseline, we use BERT and RoBERTa (Robustly optimized BERT Approach), a transformer model that was designed to improve the performance of BERT. By training with larger batches and more data, removing the next sentence pretraining objective, and dynamically masking tokens differently at each epoch, RoBERTa manages to improve on the masked language modeling objective of BERT. While both BERT and RoBERTa are primarily used for language modeling, token and sentence classification, we have implemented them to serve as our baselines to see if language models can provide human-readable summaries. BERT and RoBERTa have a maximum sequence length of 512 tokens and we used an encoder-decoder model for both transformers. For our BERT implementation, we used a BERT-initialized encoder and paired it with a BERT-initialized decoder. The same approach was used for the RoBERTa encoder-decoder implementation.

BERT (bert-extractive-summarizer). (Miller 2019) To overcome the drawbacks of the masked language models provided by BERT/RoBERTa, we used the bert-extractive-summarizer library for our extractive baseline. Similar to the BERT implementation above, the library is an encoder-only model, taking input texts and encoding them into a vector representation useful for downstream tasks. However, the BERT sentence embeddings are clustered using K-means, so that sentences with similar semantic meaning are clustered together. BERT-extractive-summarizer determines the most relevant sentences by selecting embeddings closest to the cluster centroids of the sentences it processes. The outcome would be an extractive summary representative of the document. In our implementation, we determined the number of clusters according to the number of sentences in each article—such that cluster size equals to 0.2*n, where n is the number of sentences that contains at least 60 characters in length. No fine-tuning is required for this extractive approach since the use of K-Means is unsupervised.

BART. (Lewis et al. 2019) We use the BART transformer model, a denoising autoencoder built with a sequence-to-sequence model, as another abstractive approach. It utilizes an encoder-decoder architecture that was pre-trained with corrupted text using a noising function in order to learn a model to reconstruct the original text. BART's encoder component is similar to BERT in that it learns bidirectional dependencies in text. The decoder component is similar to GPT that can auto-regressively predict tokens, allowing for text generation.

We use the sshleifer/distilbart-cnn-12-1 model that was pre-trained on the CNN/DailyMail dataset[3], made available on the Huggingface library. For our baseline approaches, we fed BART with the original concatenated articles from the Multi-News test dataset and measured performance against each article set's golden summary. We also performed a secondary baseline where we separated each topic set's articles and trained each of them against their respective golden summary, followed by evaluation against the test set. In both of the aforementioned approaches, BART automatically performs truncation on input text to fit a specified context window of 512 tokens. Decoding is performed using beam search to generate text with a maximum sequence length of 142 tokens.

Longformer-Encoder-Decoder. (Beltagy, Peters, and Cohan 2020) As a comparison to our two-stage models, we utilize a Longformer-Encoder-Decoder (LED) model from the Huggingface transformers library

[3]https://github.com/abisee/cnn-dailymail

to perform one stage summarization, where input text includes the <story_separator_special_tag>. We utilize LED-base, which has 6 layer encoder and decoder stacks that mirror the architecture of BART-base. The model uses parameters which are initialized from BART, and the position embeddings are extended to 16,000 tokens to accommodate longer input documents. The modified self-attention mechanism (windowed local-context self-attention with a task-motivated global attention) reduces time and memory complexity to scale linearly with input length. LED was finetuned using a maximum input length of 4,096 tokens and trained to generate summaries up to a maximum of 512 tokens.

Ensembles. For our baseline approaches, we would feed models with concatenated articles from the Multi-News training and validation sets. A downside to this approach is that since articles in a topic set are concatenated together, models like BART place more emphasis on articles that appear earlier rather than those that appear later due to the maximum sequence length limitation.

To account for the context windows that transformer models are generally limited by, we ensemble models together to create a hybrid summarization approach. Using either extractive-abstractive or abstractive-abstractive approaches, we process text in two stages. The two-stage ensemble approach distills important information from each article and decreases the context window for the final abstractive model. After implementing an abstractive or extractive model at the first stage, we concatenate these shortened articles and feed them into a second stage abstractive model to output a final summary.

For our fine-tuned ensemble approaches, we performed an explode function to map individual articles to their golden summaries. We fine-tuned BART using these individual articles from the training and validation sets (we refer to this step as Stage 1). This implementation incorporates single-document summarization since the ratio of article to summary is 1:1. After the Stage 1 BART model is fine-tuned, we use the model to make predictions on the training and validation sets to generate each article's respective summaries. These summaries are then concatenated according to their respective topics and fed into a separate BART model for fine-tuning (we refer to this step as Stage 2). Once fine-tuning is complete, the test set is tokenized and fed through the Stage 1 model, concatenated, and processed by the Stage 2 model. These final predictions by the Stage 2 BART model are then evaluated against the golden summaries. To accommodate for the longer text sequences via concatenation, we specify a maximum sequence length of 768 tokens for our ensemble BART models. This approach is similar to the other ensemble combination we implemented, BERT-BART, which uses BERT-extractive-summarizer in Phase 1 and BART in Phase 2.

5. Evaluation

To evaluate the performance of our models, we will utilize ROUGE (Recall-Oriented Understudy for Gisting Evaluation), a commonly used metric in machine translation and document summarization tasks developed by Lin (2004). ROUGE measures the similarity between prediction and reference summary by N-gram co-occurrence. Common ROUGE variants are ROUGE-1 (measures unigram overlap), ROUGE-2 (measures bigram overlap), and ROUGE-L (measures the similarity of the two text sequences in the sentence-level using the longest common subsequence between the prediction and reference summary). While we report the recall, precision, and F1-scores of the three ROUGE variants, our primary evaluation metric is ROUGE-2 F1-Score since we found that summaries with higher ROUGE-2 scores tend to be more readable and topic-relevant.

6. Analysis

Baseline Results

As expected, the fine-tuned BERT language models did not provide adequate summaries and yielded low ROUGE-2 scores. While their generated summaries were not directly relevant to the articles that they processed, they were still thematically similar to the ground truth. As illustrated by Appendix A, in both baselines (pre-explode and post-explode), BERT and RoBERTa were still able to create summaries about murders, even though the specific details were all incorrect. Due to their poor performance, we opted not to

implement BERT or RoBERTa in an ensemble since their generated summaries would not be appropriate inputs for the abstractive stage.

The BERT-extractive-summarizer performed well as it provided the relevant details to the articles it processed. However, there were details and phrases that were repeated throughout the extract, likely attributed to the multiple articles that it was fed. Through K-Means clustering, embeddings of sentences that convey similar information from different articles would likely be closest to the cluster centroids and thus, become selected as the most relevant sentences for summarization. Furthermore, the <story_separator_special_tag> token that separated articles was also included in the summaries. This issue was resolved after separating articles, removing the tokens, and treating the task as single-document summarization. After the explode step, extracted summaries generated for individual articles within the same topic were more succinct than those in the previous baseline. Furthermore, the summaries provided supporting and complementary information so that no article returned an exact duplicate summary, but was still relevant to the golden summary. However, summaries had punctuation issues, which could be the result of the unsupervised learning process and lack of fine-tuning performed. It is important to note that after the explode step, ROUGE-2 F1-scores for the extractive summarizer decreased relative to its previous baseline from 11.62 to 8.25, mainly because each article generated within the same topic only had a "piece of the puzzle" when compared to the golden summary. Nevertheless, summaries were still relevant and provided the main ideas for their respective topics. After merging the extracted summaries into one large summary, ROUGE-2 F1-score improved to 12.02. This behavior points to the limitations of using ROUGE as our metric, since it focuses on N-gram overlap with human-written golden summaries instead of topic/idea relevance. Individual extracted excerpts may not be compatible with the summaries in terms of exact wording and vocabulary, but we found that the excerpts were still topically relevant.

The BART abstractive model provided very good baseline results and readable summaries that were in the similar vernacular as the training data. For example, the first sentence generated provided the most pertinent information, similar to the format that legitimate news articles utilize. Other benefits of the model we noticed was that it reorders and recombines sequences of words and sentences to make summaries more readable. BART took phrases that appear later in the training data and put them at the forefront to create new complex sentences. However, one drawback we saw for the BART model was reminiscent of the BERT model, in that it would generate new words that did not originally appear in the training data, but were still thematically relevant, such as substituting "half-sister" for "half-brother." Furthermore, due to the maximum encoder length, details from the articles at the forefront of the concatenated training data were emphasized in the generated abstracts. We hypothesized that the explode step would eliminate this problem. Interestingly, however, the BART model performed worse after article separation than it did in the previous baseline from 12.94 to 12.23, but still yielded relevant and readable summaries. This may be attributed to the performance of BART's encoder model, whose language model adds or substitutes thematically relevant words but are not directly relevant to the golden summary. Generated summaries became factually inconsistent with their respective golden summaries. Similar to the behavior observed with the summaries generated by BERT-extractive-summarizer, when individual BART abstracts were merged according to their respective topics, the large summaries yielded a significant increase in ROUGE-2 F1-Score to 17.15.

The LED model performed well at producing complete and coherent text. The ROUGE-2 scores are generally higher than other models, and the text summaries are among the most human readable and similar in style to summaries produced manually. The ROUGE scores may be inflated due to LED producing longer output sequences than other models since it was limited to a maximum of 512 tokens. As illustrated in Appendix A, LED summaries incorporate multiple news outlets based on the input text given. However, these references appear to be caused by the model concatenating several abstracts into one summary—there are individual chunks of abstracts pieced sequentially. Similar to BART, the LED model paraphrases text from the input documents as phrases are drawn from the entire document set, but sentences can still closely resemble the source material. Furthermore, there are notable instances of LED inferring new vocabulary that isn't contained in the input text. In the Appendix A example, the model lists newspapers as sources that are not cited in the original text. Since LED parameters were initialized with BART parameters and the model incorporates BART's exact architecture in terms of number of layers and hidden sizes, this explains why we see some behavioral similarities with our BART implementations.

Ensemble Results

We had hypothesized that coupling two abstractive models was akin to "playing telephone", would introduce some noise, and obscure details for the final generated summary. Surprisingly, our ensemble approach of BART-BART yielded higher ROUGE-2 F1-Scores than previous BART implementations. On the other hand, the extractive-abstractive approach that was thought to follow human intuition to multi-document summarization, performed worse than any of the previous BART or BERT-ext-summ models, as it had mistook key details in its summaries. It is possible that the lack of fine-tuning in the Phase 1 extractive step led to diminished returns in performance during the Phase 2 fine-tuned abstractive step.

Upon inspection, the BART-BART summaries included relevant information in the opening lines, but would slowly degrade in coherence and in coreference resolution. We believe that the reason for the model's higher ROUGE-2 F1-Score is due to its ability to infer vocabulary similar to that of the golden summaries. For example, as shown in Appendix A, the Chicago Sun-Times newspaper was referenced by name in the BART-BART abstractive summary, even though the newspaper was not present in the training data, but was referenced in the golden summary. Being able to infer vocabulary by coincidence yielded a higher ROUGE score. However, similar to the previous BART implementation, the model does manage to paraphrase and connect sentences together that are located in separate parts of the input texts. The model continues to show the ability to take information far apart from each other and make connections between them.

7. Limitations and Future Work

Due to computational and time limitations, every model we implemented—except for BERT-ext-summ—was fine-tuned for 1 epoch. It is likely that more training time would have led to better performance since training and validation cross-entropy loss calculated during 1 epoch of fine-tuning consistently decreased throughout the training process. In addition, training times using the ensemble approach is costly since there are multiple steps to build the pipeline and fine tuning on the exploded dataset creates additional training examples. We needed to:

1. Encode training and validation data using a tokenizer.
2. Fine-tune a BART model on the training and validation data.
3. Generate, decode, and output Phase 1 summaries for the training and validation data.
4. Concatenate Phase 1 summaries according to their topic.
5. Fine-tune a second BART model using the concatenated summaries.
6. Process the test data using the above steps by tokenizing, generating, decoding with the first BART model, followed by concatenating the output summaries according to their topic.
7. Use the second BART model to generate and decode summaries based on newly concatenated test summaries.
8. Evaluate the second BART model against the test set's golden summaries.

Another weakness in our ensemble is that the Phase 1 and Phase 2 models are fine-tuned separately. Thus, the Phase 2 model does not provide performance feedback to the Phase 1 model and Phase 1 training does not necessarily improve Phase 2 performance. Instead, we believe that having a unified language model where back-propagation would update the two models during the same training phase would potentially provide better summarization results, especially in our extractive-abstractive approach. We believe that the lack of fine-tuning in the BERT-extractive-summarizer drove poor performance in its ensemble implementation.

The LED model's max input token length was set to 4,096 tokens to accommodate the vast majority of training examples without truncating text. Although the LED model was chosen as a comparison for its linear time complexity, this still caused significant training and generation times.

Our BART model summaries were limited by a max_length parameter of 142 tokens during generation, which could cause summaries to abruptly end mid-sentence. Each topic set varies in both article count and word count. For example, a maximum sequence length of 142 might be too limiting for a topic set

with five articles of medium to long length. Thus, more work should be done to allow for a more dynamic approach to specifying a maximum sequence length according to topic. Another possible reason for summaries stopping abruptly is due to the context window of our transformer models. During the tokenization process, articles are truncated to fit into an encoder length 512 tokens so the context window can stop mid-sentence. Furthermore, our flat concatenation process naively merges generated articles in Phase 2, so that summaries can be merged together mid-sentence. Future work can consider hierarchical concatenation as an alternative, which can preserve cross-document relationships and obtain semantic-rich article representation for better model learning (Ma et al. 2020).

We also saw our abstractive models infer new vocabulary that was not present in the training data. These "hallucinations" turn out to be a common occurrence in abstractive summarization. Future work could adopt the Herman system developed by Zhao, Cohen, and Webber (2020), a system that scores candidate summaries during generation via beam search and verifies whether entities and quantities are present in the training data.

The Multi-News dataset is largely unbalanced in that the large majority of topics only have two articles. The dataset is not representative of multi-document summarization scenarios in the real world, where users are more likely to request more than two articles to summarize into an abstract. Furthermore, it is difficult for us to gauge how our results compare to those of other papers that utilized the Multi-News dataset, since other papers have not reported the erroneous web-crawler generated artifacts that are found throughout the training data.

Finally, we recognize that while they are convenient as an automated metric, ROUGE scores are not perfect measures of summarization since they merely measure N-gram overlap between predictions and ground truth. Due to the vast vocabulary of the English language, generated summaries can still convey similar or equivalent meanings even if they do not share the same words as the golden summaries. Furthermore, ROUGE is not able to determine fluency, as illustrated by our examples of generated summaries by the various models. It can be argued that the BART summaries were more coherent and comprehensible than the BART-BART summary. Nevertheless, summaries from both models still had issues with completeness and correctness. In the end, since the Multi-News golden summaries are written by humans with idiosyncratic writing styles, as of yet, there is no ideal automated way to evaluate the informativeness, coherence, and readability of summaries. Human evaluation remains the standard.

8. Conclusion

In this paper, we present various model implementations using transformers to perform abstractive summarization. In terms of ROUGE score, we found that a hybrid approach of abstractive-abstractive (BART-BART) yielded the best results of any of our full self-attention implementations. However, we still observed issues concerning the quality of such generated summaries. The BART-BART model slightly underperformed relative to LED, which was purpose-built to process longer input sequences. Nevertheless, both models shared similar drawbacks in summary generation, namely inferring new vocabulary not present in the training data. Due to computational restraints, we were limited to 1 epoch of fine-tuning for smaller models of BART. We believe that a combination of a larger model (such as bart-large-cnn or Pegasus) and fine-tuning with more epochs would increase the performance of our self-attention abstractive models, as well as their downstream ensemble models.

References

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. 2020. "Longformer: The Long-Document Transformer," April. https://arxiv.org/pdf/2004.05150.pdf.

Di Fabbrizio, Giuseppe, Amanda Stent, and Robert Gaizauskas. 2014. "A Hybrid Approach to Multi-Document Summarization of Opinions in Reviews." In *Proceedings of the 8th International Natural Language Generation Conference (Inlg)*, 54–63.

Fabbri, Alexander R., Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. "Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model," June. https://arxiv.org/pdf/1906.01749.pdf.

Hokamp, Chris, Demian Gholipour Ghalandari, Nghia The Pham, and John Glover. 2020. "DynE: Dynamic Ensemble Decoding for Multi-Document Summarization," June. https://arxiv.org/pdf/2006.08748.pdf.

Lewis, Mike, 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," October. https://arxiv.org/pdf/1910.13461.pdf.

Lin, Chin-Yew. 2004. "ROUGE: A Package for Automatic Evaluation of Summaries." In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics. https://aclanthology.org/W04-1013.

Liu, Peter J., Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. "Generating Wikipedia by Summarizing Long Sequences," January. https://arxiv.org/pdf/1801.10198.pdf.

Liu, Yinhan, 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," July. https://arxiv.org/pdf/1907.11692.pdf.

Ma, Congbo, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2020. "Multi-Document Summarization via Deep Learning Techniques: A Survey," November. https://arxiv.org/pdf/2011.04843.pdf.

Miller, Derek. 2019. "Leveraging Bert for Extractive Text Summarization on Lectures," June. https://arxiv.org/pdf/1906.04165.pdf.

Subramanian, Sandeep, Raymond Li, Jonathan Pilault, and Christopher Pal. 2019. "On Extractive and Abstractive Neural Document Summarization with Transformer Language Models," September. https://arxiv.org/pdf/1909.03186.pdf.

Zaheer, Manzil, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, et al. 2020. "Big Bird: Transformers for Longer Sequences." *Neural Information Processing Systems (NeurIPS) 2020*, July. https://arxiv.org/pdf/2007.14062.pdf.

Zhao, Zheng, Shay B. Cohen, and Bonnie Webber. 2020. "Reducing Quantity Hallucinations in Abstractive Summarization," September. https://arxiv.org/pdf/2009.13312.pdf.

Appendix A. Qualitative Results

We present the following generated summaries from various model implementations to illustrate differences in summary quality.

Golden Summary:

“– A shocking story out of a Chicago suburb: Prosecutors say a 14-year-old girl set her alarm for 7:50am Tuesday, woke up, and went downstairs to the kitchen as any normal teenager might do. Except that in this case, police say the girl grabbed a 4-inch long knife, went back upstairs, and stabbed her sleeping 11-year-old half-sister to death—more than 30 times in the neck, chest, and hands, NBC News and the Chicago Sun Times report. Her apparent motive is perhaps even more shocking: With every stab, prosecutors say she told the victim the younger girl was ungrateful for the older girl’s housework; the victim had also punched her sister the night before, authorities say. They allege the 14-year-old then showered to wash off the blood before calling 911, telling police a Hispanic male intruder had entered the home the girls shared with their mother—who wasn’t home at the time—and did the deed, CNN reports. That led to the lockdown of three nearby schools before the girl’s story unraveled. When detectives pointed to strands of hair tying her to the crime, police say she admitted to it, adding that she had contemplated the killing the night before. The girl now faces first-degree murder charges in juvenile court; a decision on whether to move the case to adult court will be made in a few weeks.”

Table 1: Modeled Summaries

Model	Generated Summary
BERT - Baseline with original training set	‘– it’s not clear how many people have been killed in the last two years, but the ap reports that the number of people in the country has been more than 1, 000, according to a new report from the ap. the ap notes that in the first quarter of the year, more than 30, 000 people were injured in the violence in the city of st. louis, where the victims were taken to the hospital. " we’re going to have to do it, " says one of the victims. " i’m not sure what happened to me, " said a police officer. " it’ s the the ap adds, per the new york times. " the’
BERT - Baseline with exploded training set	“– police in chicago say a 13 - year - old girl was stabbed to death in a home in the middle of the night, reports the chicago sun - times. police say the girl, who was stabbed in the head by her sister, was stabbed at the home of her sister’s sister, who has been charged with murder. the girl’s mother, who had been in the home since the attack, tells the tribune that she was stabbed by a young girl who was trying to kill her sister. she says the girl had been stabbed to the head, and she says she had been beaten to death by her brother. police say a neighbor tells the chicago tribune that her sister and her”
RoBERTa - Baseline with original training set	‘– A 13-year-old girl was killed in a home invasion last week, police say. Police say the girl’s mother and her two brothers, ages 14 and 17, were found dead at a home in St. Paul, Minnesota, when they went to check on their 4-month-old daughter, who had been taken to a hospital for treatment, reports the Minneapolis Star Tribune. The girl is believed to have been shot by her parents, but police believe she may have been hit by a car or thrown into an apartment building. “She was very young,” says a police spokesman. "It’s not clear what’

Model	Generated Summary
RoBERTa - Baseline with exploded training set	' The death of a 13-year-old girl in California has led to more than a dozen arrests in the case, the Los Angeles Times reports. According to CBS Los Angeles, the girl, identified only by her last name, was found dead in her bedroom at her parents' home. Police say the girl was playing with a Nintendo DS when she was hit by a baseball bat. She was pronounced dead at the scene. The girl's mother, who was in her 20s, says she was "very upset" when she saw the bat hit her daughter. "She was like, 'Oh, my God, this is'
BERT-ext-summ - Baseline with original training set	'story highlights the 14-year-old called 911, police said officers found her sister upstairs stabbed 40 times neighbors say the sisters seemed close and played together the 14-year-old is in a juvenile detention center an 11-year-old illinois girl was stabbed 40 times by her 14-year-old sister, police say, and the reason why has many in the small chicago-area community of mundelein reeling in disbelief. pray, first of all, for the family that has been devastated by this." intruder in the house police received a 911 call about the incident at 8:30 a.m tuesday. the teen said she cooked dinner for her younger sibling and performed other chores. "they would always be together," mary ann gryder, a neighbor told the affiliate. prosecutors said they'll decide in the coming weeks whether to charge her as an adult. her family, meanwhile, left the juvenile courthouse without commenting, escorted by security guards. <story_separator_special_tag> community struggles to come to grips with a 14-year-old girl accused of stabbing her 11-year-old half sister to death. (during a wednesday court hearing, prosecutors said the 14-year-old girl set her alarm early, went downstairs to the kitchen and got a knife. officials say a decision will be made in a few weeks whether to move the case to adult court. when detectives told her later in the afternoon they had strands of hair that would eventually point to the real suspect, the girl admitted she did it, prosecutors said. guenther said the suspect is the one who called 911 to report the emergency and answered the door when police arrived four minutes later. he said whatever transpired happened shortly before the 911 call was placed. the discovery of the girl's body prompted a lockdown of nearby schools. while both girls attended the same school, guenther declined to identify it.'
BERT-ext-summ - Baseline with exploded training set (3 summaries)	'story highlights the 14-year-old called 911, police said officers found her sister upstairs stabbed 40 times neighbors say the sisters seemed close and played together the 14-year-old is in a juvenile detention center an 11-year-old illinois girl was stabbed 40 times by her 14-year-old sister, police say, and the reason why has many in the small chicago-area community of mundelein reeling in disbelief. there also is an effort to have residents leave their lights on friday night in remembrance. the teen said she cooked dinner for her younger sibling and performed other chores. "they would always be together," mary ann gryder, a neighbor told the affiliate.'

Model	Generated Summary
	‘girl, 14, charged with stabbing half-sister, 11, more than 30 times by jon seidel and ruth fuller staff reporters article extras updated: the flush-faced 14-year-old mundelein girl sat with her palms flat on the table in the lake county juvenile courtroom, listening carefully. the woman appeared stunned, but finally broke down in tears, as she heard the account of her younger daughter’s final moments. prosecutors said they’ll decide in the coming weeks whether to charge her as an adult. the discovery of the younger girl’s dead body led briefly to a lockdown tuesday of nearby schools after the older girl allegedly told police a hispanic male broke in, stabbed her half-sister and fled. but she later confessed when confronted with hair strands found in her half-sister’s hands — which could be tested for dna — and revealed an even more troubling narrative that has shaken her far north suburban home.’ ‘community struggles to come to grips with a 14-year-old girl accused of stabbing her 11-year-old half sister to death. (prosecutors said that after the stabbing, the 14-year-old girl called police and told them there was a hispanic male intruder stabbing her little sister. when detectives told her later in the afternoon they had strands of hair that would eventually point to the real suspect, the girl admitted she did it, prosecutors said. guenther said the suspect is the one who called 911 to report the emergency and answered the door when police arrived four minutes later. while both girls attended the same school, guenther declined to identify it.’
BART - Baseline with original training set	‘– A 14-year-old Illinois girl was stabbed 40 times by her sister, reports the Chicago Tribune. Police say they found an intruder with a kitchen knife at 8:30am Tuesday morning, and police found her unconscious in the home of Mundelein, allegedly stabbing her half-brother, who called 911 to be treated for medical attention. “This incident is a heartbreaking tragedy that defies understanding,” per WLS reports. The teen told emergency dispatchers that she thought she was ungrateful. She later admitted to the attack on her sister’s birthday. "I am asking the Chicago area, please pray for’
BART - Baseline with exploded training set (3 summaries)	‘– A 14-year-old Illinois girl was stabbed 40 times by her own sister in the Chicago suburb of Mundelein, Ill., on Tuesday morning, and police say. The teen told police that an intruder had barged into the home with a kitchen knife after she allegedly attacked her half sister, reports WLS. She later died, and then went downstairs and grabbed a knife, and took a shower to wash off the blood, CNN reports. Police say the teen’s brother confessed to the attack, killing her half-brother, who tried to kill her. “This incident is a heartbreaking tragedy that defies understanding,” ’ ‘– A 14-year-old girl allegedly stabbed her half-sister in the Chicago suburb of Mundelein, Ill., has been charged with first-degree murder. According to court documents obtained by a judge on Wednesday, prosecutors say that the teen’s mother “didn’t show enough gratitude for all she had done everything she made the girl dinner six times,” the Chicago Tribune reports. She told police that she’d killed her 11-year old brother, who was stabbed more than 30 times. The older sister, whose body was found in a garage door to look at the boy’s body, and then fled.’

Model	Generated Summary
	'– A 14-year-old girl stabbed to death in her half-sister because she didn't appreciate the things she had been doing around the house, reports the Chicago Sun-Times. Police say the teen was stabbed 30 times in the neck, chest, and arms. She went downstairs to the kitchen and got a knife, and then returned upstairs to her sister's bedroom, police say. The older girl said she had cooked dinner six times that she'd been hitting her while telling her how she didn't appreciate her things the older girl had been done around. "I just want to ask the coroner's office'
BART-BART Ensemble with exploded training set and subsequent concatenation Phase 1 and Phase 2	'– A 14-year-old girl stabbed her half-sister because she didn't appreciate the things she had been doing around the house, reports WLS. According to court documents obtained by a judge on Wednesday in Mundelein, Ill., police say that an intruder had barged into the home with a kitchen knife, and arms were stabbed 40 times in the neck, chest, per the Chicago Sun-Times. "I just want to ask the coroner's office for all she made the girl dinner six times," according to court records. The older sister, whose body was found in a garage door to wash off the'
BERT-ext-summ-BART Ensemble with exploded training set and subsequent concatenation Phase 1 and Phase 2	'– A 14-year-old girl was stabbed 40 times by her own sister in the Chicago suburb of Mundelein, Illinois, and police say. The suspect is accused of stabbing her half-sister confessed to killing her 11-month-old sister, reports the Chicago Tribune. Police say they found a Hispanic male intruder at the same time she called 911 to report the AP. "It's not clear what happened," per the Chicago Sun-Times. "I think it was going to be together," says a neighbor. She told police that there was a Mexican man who allegedly stabbed her younger sister to death with her'
LED	'– An 11-year-old Illinois girl was stabbed 40 times in an upstairs bedroom of her sister's home Tuesday morning, the Chicago Sun-Times reports. Police found the girl unconscious and in need of medical attention in the upstairs bedroom, and she later died. Police say the girl, who called 911, initially told officers an intruder had entered the home, stabbed her sister 40 times, and then fled, CNN reports. The teen, who initially told police she was mad that her sister didn't appreciate all she did for her, was arrested and charged with murder. The Chicago Tribune reports that police say the teen told them an intruder entered the house and stabbed the victim 40 times with a kitchen knife because the teen thought she was ungrateful. The girl's sister, who also called police, initially said an intruder was to blame for the attack, but police say she later admitted to the attack. "This incident is a heartbreaking tragedy that defies understanding," said Mundelein Mayor Steve Lentz. "I am asking the Chicagoland area, please pray for us. pray, first of all, for the family that has been devastated by this." A group called "mundelein moms" is raising money for the girl's funeral and said it raised $6,381 in just one day.'

Appendix B. Quantitative Results

Table 2: Baseline - Concatenated Inputs with Separator Token

	R1-P	R1-R	R1-F1	R2-P	R2-R	R2-F1	RL-P	RL-R	RL-F1
BERT	41.52	21.57	27.74	6.66	3.40	4.40	22.28	11.69	14.96
roBERTa	43.95	21.81	28.47	7.16	3.46	4.56	20.94	10.41	13.57
BART	**56.59**	27.81	36.39	**20.18**	9.87	12.94	**29.11**	14.35	18.75
BERT-ext-summ	37.36	**46.3**	38.41	11.54	**13.79**	11.62	17.34	**20.93**	17.52
LED	49.50	36.54	**40.58**	17.34	12.80	**14.22**	25.44	18.75	**20.81**

Table 3: Baseline - Exploded Individual Articles (Phase 1 only)

	R1-P	R1-R	R1-F1	R2-P	R2-R	R2-F1	RL-P	RL-R	RL-F1
BERT	48.28	24.21	31.49	10.13	5.02	6.55	24.42	12.32	15.97
roBERTa	41.37	20.1	26.43	6.34	2.99	3.98	21.89	10.67	14.01
BART	**56.32**	26.19	34.91	**19.82**	9.15	12.23	**29.56**	13.77	18.34
BART (merged summaries)	43.28	**51.54**	**45.71**	16.24	**19.34**	**17.15**	19.68	**23.60**	**20.84**
BERT-ext-summ	47.45	25.69	30.45	13.51	6.76	8.25	25.38	12.86	15.55
BERT-ext-summ (merged summaries)	35.53	49.75	38.81	11.11	15.28	12.02	16.1	22.27	17.43

Table 4: Complete Ensemble (Phase 1 and 2)

	R1-P	R1-R	R1-F1	R2-P	R2-R	R2-F1	RL-P	RL-R	RL-F1
BART/BART	**57.26**	**27.81**	**36.54**	**21.71**	**10.44**	**13.77**	**30.3**	**14.74**	**19.35**
BERT-ext-summ/BART	54.83	26.92	35.24	16.81	8.15	10.72	26.92	13.24	17.31

Appendix C. Model Architecture

Figure 1 shows the pipeline architecture. Articles are separated and paired with their golden summary using an explode function, then fed into a Phase 1 model which produces individual summaries. These summaries are then concatenated according to their original topic to form one large summary. This intermediate summary is then fed into a Phase 2 model to produce a final abstractive summary.

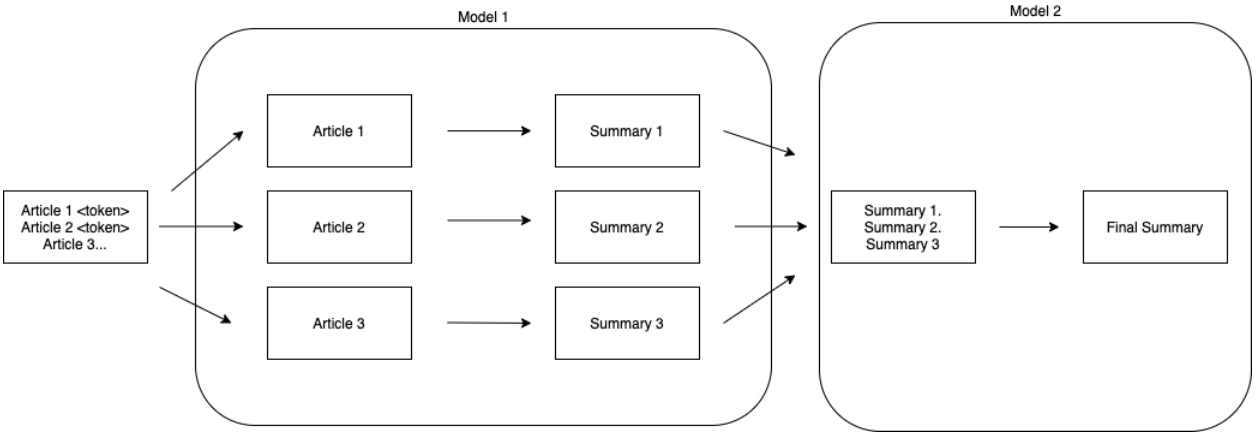

Figure 1: Two-Stage Model Architecture

Appendix D. Muti-News Errors and Text Distribution

Figure 2 shows the distribution of the number of articles given per summary in the Multi-News dataset as well as the number of words per article and summary. Article-summary sets are heavily weighted towards fewer articles with 2 being the most common. The average golden summary is generally shorter than one of its input articles.

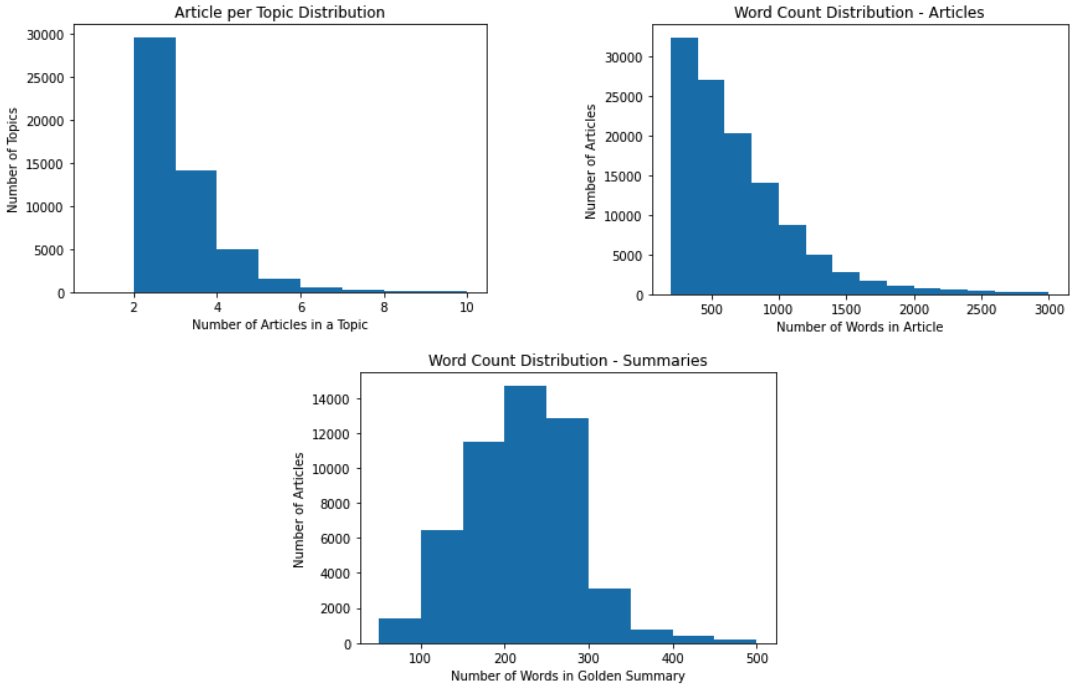

Figure 2: Multi-News Text Distribution

Examples of error articles removed from Multi-news dataset:

- 'these crawls are part of an effort to archive pages as they are created and archive the pages that they refer to. that way, as the pages that are referenced are changed or taken from the web, a link to the version that was live when the page was written will be preserved.then the internet archive hopes that references to these archived pages will be put in place of a link that would be otherwise be broken, or a companion link to allow people to see what was originally intended by a page's authors.the goal is to fix all broken links on the web . crawls of supported "no more 404" sites.'
- 'the seed for this crawl was a list of every host in the wayback machine this crawl was run at a level 1 (urls including their embeds, plus the urls of all outbound links including their embeds) the warc files associated with this crawl are not currently available to the general public.'
- 'crawl of outlinks from wikipedia.org started march, 2016. these files are currently not publicly accessible. properties of this collection. it has been several years since the last time we did this. for this collection, several things were done: 1. turned off duplicate detection. this collection will be complete, as there is a good chance we will share the data, and sharing data with pointers to random other collections, is a complex problem. 2. for the first time, did all the different wikis. the original runs were just against the enwiki. this one, the seed list was built from all 865 collections.'

Appendix E. Generated Paper Summaries

The summaries contained in the table below were generated on the main body of this paper and did not include the abstract. Although the models were trained on multiple news articles and not single academic papers, the length and structure of the paper provides an interesting opportunity to test the models' functionality.

Table 5: Paper Summaries

Model	Summary
Out-of-the-box BART	' The task of providing summaries for the most important information to the world, is being explored by researchers in the University of New York (BERT) and a number of documents that can be used to summarise more than 2-1.
Baseline BART	'– A new study has found that a way to make it easier for people to read about the most important information on the subject of an online study in the Journal of Science. The New York Times reports that "simply" can be used to summarize what they describe as "the most important pieces of information," says researchers at the University of Cambridge, who have worked out with the word "tremarkable" and "unprecedented" number of documents: "It's difficult to understand how many people are able to do so." That's not only one person would like to use the phrase "snow-to-'
Phase 1 BART	'– A new study out of the University of California has found that it's a way to make it easier for people to read: "I'm not going to do so," says researchers in the journal Proceedings of the National Academy of Sciences, which is based on computer science. The paper calls it "one of the most important pieces of information" and "extremendable summaries." That's what's being described as an "unprecedented" number of documents, per the Washington Post. Researchers say they've developed a system designed to produce more than 3,000 times more often available online, reports the New York Times'
Phase 2 BART	' A lot of work on news, reviews, emails, and emails, legal documents, and it is a sequence-to-sequence task, the volume that readers face can become overwhelming. As media becomes even more readily available online. The intuition behind this process, taking input text to return another set of real-world applications, such as generating convenient summaries from different sources and in comprehension tasks, using few phrases directly from the source text and inferring evidence to produce "the most important pieces of information," according to researchers. In this study, we will explore various ways of understanding the English language for the multi-document setting,'

Model	Summary
LED	'– When faced with three documents to perform summarization, a person would likely read each of the three documents, find the lines most relevant to the topic by extracting facts and quotes from each document, and then reword them together to assemble a final abstractive output. The results are similar to those of the previous BART implementation, but they differ in the way the model is trained. For example, the model uses an encoder-only model, taking input texts and encoding them into a vector representation useful for downstream tasks. The model is similar to BERT, but it uses a modified self-attention mechanism and does not require significant text truncation to produce abstractive summaries. The result is a higher ROUGE-2 F1-Score, which is better than the average score of BERT-Extractive-Summarizer, which has a maximum sequence length of 512 tokens. It's not clear why this is, but the researchers say it's possible that the model was designed to improve performance in the multi-documenting task. "We had hypothesized that coupling two abstractive models was akin to 'playing telephone,' but BART's encoder model, whose language model adds or substitutes thematically relevant words, would introduce some noise, and obscure details for the final generated summary," the researchers write in a paper published in the journal Proceedings of the National Academy of Sciences.'