Two-Stage Text Summarization with Pretrained Transformers

Yifan Li, Yukai Yang, Yifei Li, Tianze Zheng
School of Engineering and Applied Science
University of Pennsylvania
{yfli,yukaiy,liyifei,leozheng}@seas.upenn.edu

Abstract

Text summarization is the task of automatically generating short highlights for a long source document. The current models can basically be divided into two categories: 1) extractive, which selects sentences or words from the original article 2) abstractive, which generates new natural language sequences. Extractive models are robust but not like human-written summaries. Abstractive models are flexible but may contain factual inconsistencies.

Therefore, in our project, we developed a 2-stage model for text summarization. It combines the ability to filter informative sentences of the extractive approach with the ability to paraphrase of the abstractive approach. Our best model achieves a ROUGE-L F1 score of 39.82, which outperforms the strong baseline.

1 Introduction

Text summarization is the task of automatically generating natural language summaries from the source documents that preserve most of the salient ideas of the original. By condensing words into short summaries, summarization is useful in many downstream tasks such as generating news digests and reports.

Different from other NLP tasks, summarization requires a wide-coverage of natural language understanding beyond individual words and sentences. There are two broad approaches to summarization, *extractive* and *abstractive*. *Extractive method* retrieves summaries directly from the original sources without any changes or modifications. It is a binary classification task that labels words in each sentence whether they are included in the summary or not. *Abstractive method* generates novel words and new phrases not appeared in the sources to summarize the original text, which requires language generation capabilities. *Extractive* system is more robust and straightforward, and *abstractive* system is more flexible for different situations. Figures below are examples of *extractive* and *abstractive* methods.

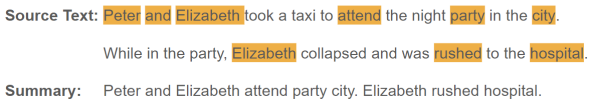

Figure 1: Extractive Methods

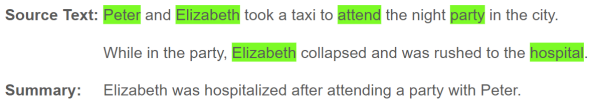

Figure 2: Abstractive Methods

Consider another example of text summarization of the data (Liu et al., 2019a):

Article:

- Jason listened to the weather and heard it was going to be sunny. He thought the kids might like to go swimming. He gathered up the swimsuits, towels and sunscreen. Jason and the kids got into the truck and drove to the beach. They spent the next 2 hours playing and splashing in the surf.

Summaries:

- Jason saw a nice weather forecast and went to the beach with his kids for 2 hours.
- Jason took the kids swimming at the beach on a sunny day.
- Jason decided to take the kids to the beach since it was a sunny day.

A short piece of story can have different summaries, each having its own focuses and details.

We found this task intriguing because it leads to many real-world applications. For example, it can be used to title a piece of news, or compose a brief summary for search engine results. A short summary will be helpful for readers to get the hint about what the whole passage about, and decide if it is worth their time to read the whole passage.

2 Literature Review

2.1 Extractive-Based

Cheng and Lapata (2016) first formulates the task of extractive summarization as a sequence labeling problem and proposes to solve it using an encoder-decoder framework. Their model includes a neural network-based hierarchical document reader and an attention-based hierarchical content extractor. Their neural sentence extraction model outperforms the LEAD and LREG baselines with a significant margin on DUC 2002 test dataset using ROUGE. However, models like this assume the independence of sentences and make binary classifications for each sentence, leading to high redundancy.

Zhou, Yang, Wei, Huang, Zhou, and Zhao (2018) addresses the issue with an auto-regressive decoder that makes the scoring of different sentences dependent on each other. In such way, sentence scoring can be aware of previously selected sentences and selection can be simplified. Their NEUSUM obtains SOTA by the time of publication on the CNN/Daily Mail dataset, scoring 41.59 on ROUGE-1 and 19.01 on ROUGE-2.

Recently, BERTSumEXT (Liu and Lapata, 2019) includes *Trigram Blocking* to further reduce redundancy. In the encoder frame, BERTSumEXT extends BERT by inserting multiple [CLS] tokens to learn the sentence representations. For extractive summarization, BERTSumEXT stacks several intersentence Transformer layers. The model outperforms all previously proposed extractive systems.

2.2 Abstractive-Based

Abstractive based text summarization requires model to have natural language generation capabilities, and transformer-based models like GPT-2 and BART are successful in recent years in performing the task.

2.2.1 GPT

Aiming to tackle multitask by one unsupervised language model, GPT-2 (Radford et al., 2019) is trained on a sufficiently large, various, and clean dataset. In addition, it uses improved Byte Pair Encoding (BPE), a middle interpolation between character and word level language modeling as the input representation. Last, transformed-based architectures and layer normalization are indispensable components of this model.

Therefore, different from the previous frameworks which express the task as a prior, GPT-2 can be flexible in many downstream tasks. Especially in the summarization performance on the CNN/Daily Mail dataset, GPT-2 successfully renders the summary, using the first 3 sentences generated after the sequential trigger word of the article `TL;DR:`. However, it comes with the confusion of some specific details, merely on par with the performance of classic neural network baselines and barely outperforms the random-3 sentences from the article, based on the popular metrics like ROUGE 1, 2, L.

2.2.2 BART

BART (Lewis et al., 2019) adopts a standard sequence-to-sequence (seq2seq) transformer architecture with a bidirectional encoder (like BERT (Devlin et al., 2019)), and a left-to-right decoder (like GPT), thereby generalizing the two pre-training schemes. Similar to BERT, BART is trained like a denoising autoencoder by corrupting documents and reconstructing the missing tokens. The pre-training process involves randomly shuffling the order of sentences and replacing random spans of text with a single mask token. Further, BART uses BPE tokenization and is trained on the same scale as the RoBERTa (Liu et al., 2019b) model. Since the corrupted inputs are only fed into the encoder while the decoder is trained on the original text in an autoregressive fashion, BART reduces the mismatch between pre-training and generation.

Due to its seq2seq structure, BART is particularly effective when finetuned for conditional generation tasks like summarization. BART reports to outperform all previous methods on the two summarization datasets, CNN/DailyMail and XSum, with gains of up to 6 ROUGE. Qualitatively, the output summaries from BART tend to be fluent, highly abstractive, and factually consistent with the original articles for the most part.

3 Approach

As introduced in the literature review, the previous methods can generally be divided into two categories: extractive and abstractive. Both methodologies, however, has certain drawbacks.

The extractive model cannot deal with the fact that human tends to paraphrase when writing summaries. The abstractive approach suffers from the problem that off-topic sentences are also fed into the model, resulting in the possibility that some output sentences are also irrelevant. Also, we found that in practice, some abstractive models have limits for number of input tokens. If the source document is longer than the limitation, a commonly adopted solution is just to cut-off at the limitation, which introduce the risk that important information is also dropped.

Our approach attempts to take advantage of both the extractive and abstractive approach, so as to make up for each others' drawbacks. In the first stage, we use an extractive model to filter out the off-topic sentences. Then we feed only the relatively important sentences into the second stage, abstractive models to further refine and paraphrase. The model structure is demonstrated in Figure 3.

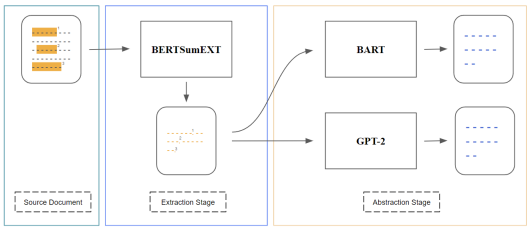

Figure 3: Model Structure

4 Experimental Design

4.1 Data

We intend to use the CNN/DailyMail dataset. The data contains more than 300k unique news articles written by journalists at CNN and the Daily Mail. The news will serve as source texts and the title will be regarded as the target summarization.

We chose this dataset because it is extensively used in recent text summarization papers. Its prevalence makes Hugging Face support it directly, so we can directly download and use the data via Hugging Face's dataset library, and use a standard train/validation/test split. The proportion of each split is 92%/4.3%/3.9%, respectively (Nallapati et al., 2016).

4.2 Evaluation Metric

As for the evaluation metrics, we use Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), determined by the difference between the generated summaries and gold summaries. ROUGE-1/2/L are chosen as our performance metrics. In detail, ROUGE-1 and ROUGE-2 refer to overlap of unigrams and bigrams between two summaries respectively. The reason we use them in conjunction is because while ROUGE-1 can capture the reuseness of single word, ROUGE-2 can indicate the fluency or readability of the summaries by the frequency of original word orderings. To put the sentence level word order into account, ROUGE-L, the measurement of the longest matching sequence of words via Longest Common Subsequence (LCS), is also implemented. Although there are other newer reference-based metrics such as BERTScore (Zhang et al., 2019) and ROUGE is limited by its inflexibility (See et al., 2017), we stick to ROUGE since it's widely used in summarization and thus bring convenience of comparison.

Mathematically, let R and C be the reference (i.e. gold summary from humans) and candidate (i.e. output summary from model), we have Equation 1 where N is the length of n-gram, s_i is sentence and $\text{Count}_{(R \text{ or } C)}(\text{n-gram})$ counts the frequency of n-gram in the reference or candidate respectively. When the denominator counts on R, the ROUGE-N is recall, otherwise it's precision. Their harmonic mean is F1 score. Similarly, we have

$$\text{ROUGE-L}(R, C) = \frac{\sum_{s_i \in R} |\text{LCS}_\cup(C, s_i)|}{\text{Count}_{(R \text{ or } C)}(\text{word})}$$

where $\text{LCS}_\cup(C, s_i)$ is the set of longest common subsequences in the candidate and single sentence from reference. $\text{Count}_{(R \text{ or } C)}(\text{word})$ is the number of words in reference and candidate respectively. Using former gives recall and using the latter gives precision. The general F1 score is computed by

$$\text{F1} = \frac{(1 + \beta^2) \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \beta^2 \cdot \text{precision}}$$

where β adjusts the relative importance of the precision and recall. Considering ROUGE favors recall, β is generally set to a high value.

$$\text{ROUGE-N}(R, C) = \frac{\sum_{s_i \in R} \sum_{\text{n-gram} \in s_i} \min\left(\text{Count}_R(\text{n-gram}), \text{Count}_C(\text{n-gram})\right)}{\sum_{s_i \in R} \sum_{\text{n-gram} \in s_i} \text{Count}_{(R \text{ or } C)}(\text{n-gram})} \quad (1)$$

In our experiment, there's just one pair of reference and candidate for each article. But generally speaking, if we have k references, the ROUGE score can be generalized by

$$\text{ROUGE}(R, C) = \max_k \{\text{ROUGE}(R_k, C)\}$$

where ROUGE can be any form of the ROUGE score.

4.3 Baselines

We implement two baselines. The weak baseline is a random word picker which renders a summary comprised by M words randomly picked from the article, where M is proportional to the length of article. Here, it's set to 30 by the average ratio of length between the article and reference in the first 100 samples. The strong baseline is called Lead-3 proposed by by See, A. *et al* (See et al., 2017), which simply concatenates the first 3 sentences as a summary. It is a popular strong baseline in the summarization domain and proved to have an outstanding performance especially on the CNN/DailyMail dataset. They explained it is possible that either the new articles tend to begin with the most important information, or extractive approaches are naturally favored by the task and ROUGE given the subjectivity of summarization and inflexibility of ROUGE. To tackle the first limitation, we introduce another strong baseline named Random-3, selecting 3 sentences randomly from the article as a summary.

5 Experimental Results

5.1 Implementation Details

BART fine-tuning. We used Huggingface's 'bart-large' checkpoint and Pytorch to implement BART fine-tuning. We first tokenized both the source articles and reference summaries using BART's BPE tokenizer. We trained the model for 18,000 steps with a warmup of 500 steps on one GPU (A100); attention dropout and label smoothing with a factor of 0.1 were also applied. Because the large model has 12 layers of encoders and decoders which barely fit into our GPU RAM, we used a small batch size of 2 and accumulated gradient every 16 steps. For generation, following (Lewis et al., 2019), we used 4 beams, limit the summary length at 142, and block duplicate trigrams to avoid repetitions.

GPT-2 fine-tuning. Likely, with the help of the pretrained GPT2LMHeadModel from Huggingface, we can fine-tune GPT-2 serving for summarization. By concatenating the original article and gold summary, we can feed GPT-2 by a tokenized input, which can turn back as the original article plus a new summary. Following lots of seq2seq training, we use cross entropy loss function to minimize the error between sequence. However, suffered from the limited number of input tokens (2014), we have to truncate the input length. Here, we truncate the article and gold summary proportionally by their ratio of length.

Two-stage training. In this setting, the extractor BERTSumEXT serves as a binary classifier that selects important sentences to include in the summary. We used off-the-shelf BERTSumEXT trained on CNN/DailyMail. Sentences with scores greater than 1.01 are selected from BERTSumEXT and fed into our abstractor (BART or GPT-2). We fine-tuned the abstractor on the extracted sentences following the same procedure described above. Note that since BERTSumEXT does not use a objective designed for two-stage summarization, we can also fine-tune BERTSumEXT on sentences with the most information coverage in our setting, which can potentially improve our two-stage performance.

5.2 Baselines

The performance on the test set is shown in Table 1. Among the metrics, ROUGE-1 only considers unigram overlap and therefore always comes with the highest scores, while ROUGE-2 is the most difficult with the lowest scores. Especially since the random-word summary is extremely unreadable, the ROUGE-2 of random word picker is close to zero. It is obvious that the random work picker is weak given its worst performance among the three. Lead-3 achieves strong performance on ROUGE-1, ROUGE-2, and ROUGE-L. This makes sense because summaries are closely related to source articles in a news dataset. Unlike

Lead-3, Random-3 gives a more moderate result, showing that it is likely a fairer strong baseline on the CNN/DailyMail, as explained in the last section.

5.3 Models

BART Table 1 shows the performance of our BART fine-tuned on CNN/DailyMail. Our reproduced BART outperforms all three baselines on all ROUGE metrics, including the strong baseline Lead-3. Notably, it even scores slightly higher than the original BART (Lewis et al., 2019) on ROUGE-1 and ROUGE-L. Our BART also beats BERTSumEXT, an extractive model, across all ROUGE metrics, indicating BART is able to generate summaries more closely matched with the reference than the extracted sentences from BERTSumEXT.

GPT-2 However, GPT-2 does not do a good job. The fine-tuned GPT-2 barely pass the weak baseline, though with a much higher fluency. Besides the aforementioned reason of truncated input, another reason is the long training time, making it hard to using the whole training set to fine-tune GPT-2 in a short term. This toy GPT-2 has been trained by just 3000 samples by 5 epochs. Therefore it trends to overfit and produce some meaningless repeated phrases.

BERTSumEXT The goal of the extractive model is to filter out sentences that are irrelevant to the topic. But the remaining sentences should still contain sufficient information for the abstractive model to further refine and paraphrase. For this reason, recall in the first-stage is much more important than precision. The original configuration of BERTSumEXT, however, constraints its output to 3 sentences because it aims to optimize the Rouge F1 directly. As a result, adopting the original version of BERTSumEXT as the extractor leads to very poor performance, no matter how the abstractor is fine-tuned.

We modified the configuration of BERTSumEXT to make it cover more information. We have tried to 1) increase the constraint of sentences. However, this does not adapt to the varying lengths of the source documents. Then we 2) pick all sentences that has higher scores than a threshold. This method proves to be more effective.

To give a quantitative comparison, table 2 the ROUGE recall score of all configurations of BERTSumEXT and an oracle reference. The reference is compiled by a modified version of the method in (Liu and Lapata, 2019) to generate the oracle: Sort all sentences in the original article, and greedily add top sentences into the tentative oracle until its ROUGE-L recall stops increasing. Based on the result, we choose 1.01 as the threshold because its recall reaches the level of the gold sequence.

Two-stage Our two-stage model combining BERTSumEXT and BART successfully outperforms running BERTSumEXT alone on all ROUGE metrics by a nontrivial margin, as shown in Table 1. We prove that with the help of abstractor, our two-stage model can paraphrase sentences from an extractive-only model to generate more readable summaries. However, it obtains lower ROUGE scores compared to our BART. We hypothesize that BERTSumEXT without fine-tuning likely do not provide sufficient information from the article for the abstractor; this can also be observed from a higher fine-tuning loss compared to BART after 18,000 steps. We can draw the same conclusion of GPT-2's two-stage model. Limited by the input length, an extractor which catches essential information first can definitely improve the quality of input sentences, making it more efficient to fine-tune GPT-2. Compared to the single GPT-2 model, the two-stage pipeline improves the ROUGE scores.

To test the performance upper bound of the two-stage approach, we combine BART with the oracle, which guarantees to output sentences with enough information to generate the reference summaries. This effectively assumes the extractor always outputs perfect sentences. Note that this is not a practical model because the oracle also needs reference summaries at test time. We fine-tune BART on the oracle-generated sentences from the training set. The model achieves very competitive ROUGE-1 and ROUGE-L F1 that are above 50, which are significantly high than BART. This indicates the potential of the two-stage approach if we further fine-tune BERTSumEXT using the oracle as the groundtruth sentences. Note that the sentences from the oracle alone produce low F1 scores on the test set since it is optimized on ROUGE-L recall.

5.4 Error Analysis

A qualitative analysis of the models is shown in Table 3 and Table 4.

BART generates highly fluent summaries and has learned to both copy from the original article

Model	ROUGE		
	1	2	L
Random Word Picker (weak baseline)	17.22	0.49	14.50
Random-3 (strong baseline #1)	28.43	8.26	25.48
Lead-3 (strong baseline #2)	40.06	17.48	36.37
Extractive			
BERTSumEXT	41.22	18.42	37.57
Abstractive			
GPT-2 (few-shot)	21.38	5.5	15.52
BART (Lewis et al., 2019)	44.16	21.28	40.90
BART (ours)	44.28	21.18	41.39
Two-Stage			
BERTSumEXT + GPT-2 (fine-tuned)	30.86	10.80	28.88
BERTSumEXT + BART (fine-tuned)	42.8	19.73	39.82
Oracle + BART (fine-tuned)	54.38	30.86	51.76

Table 1: ROUGE F1 results on CNN/DailyMail test set.

Configuration	ROUGE		
	1	2	L
Constrain Length of Sentence			
3 sentences (original)	52.3	23.4	47.6
4 sentences	59.5	27.3	54.7
5 sentences	64.5	30.1	59.7
6 sentences	68.2	32.3	63.5
Sentence Score Threshold			
Score-1.01	79.8	39.6	75.6
Score-1.03	75.9	37.3	71.5
Score-1.05	70.7	34.2	66.1
Oracle			
Gold Reference	79.0	43.5	75.2

Table 2: ROUGE recall of different configurations of BERTSumExt on test set.

and occasionally generate novel words. However, for the second article, BART confuses the player and the team that scored the goal, which is factually incorrect. It also generates a sentence irrelevant to the article at the end. Combining BERTSumEXT with BART seems to eliminates the inconsistency for the second article. Further, BART outputs sentences in a random order that appear a little unrelated to each other, while the two-stage model seems to follow a more logical order and is thus more readable. For example, in article 3, BART first generates a sentence about the casualty from a tornado and then goes on to talk about the scale of the tornado, while the two-stage model describes the tornado first before giving out the consequences. The two-stage model covers more information by including almost all locations affected by the tornado attacks mentioned in the article, while BART only considers Illinois. Despite being abstractive, BART and its two-stage version tend to copy from the article or only paraphrase a little. This makes sense due to the nature of CNN/DailyMail; a more abstractive dataset will likely give different results.

GPT-2 and GPT-2's two-stage model always suffer from the repeating sentence problem. It is likely either overfitting due to the small subsampled dataset, or caused by the wrong setting in the fine-tuning process.

6 Conclusions

In this work, we present a two-stage summarization model using pretrained transformers, concatenating an extractive model BERTSumEXT and an abstractive model, GPT-2 or BART. By feeding the abstractor the extracted key information, this method can mitigate the disadvantages of both approaches and in turn make the summaries more readable. It also eliminates the need to truncate sentences for the abstractive model due to the maximum token limit. We prove that a two-stage model can in some cases outperform a single model and has the potential to be generalized to other fine-tuning tasks.

Acknowledgments

We are grateful to Dr. Yatskar for his project suggestion and teaching assistants in CIS 530 especially Alyssa Hwang for their fruitful comments, corrections and inspiration.

References

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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

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

Peter J. Liu, Yu-An Chung, and Jie Ren. 2019a. Summae: Zero-shot abstractive text summarization using length-agnostic auto-encoders. *ArXiv*, abs/1910.00998.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

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

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences.

Supplemental Materials

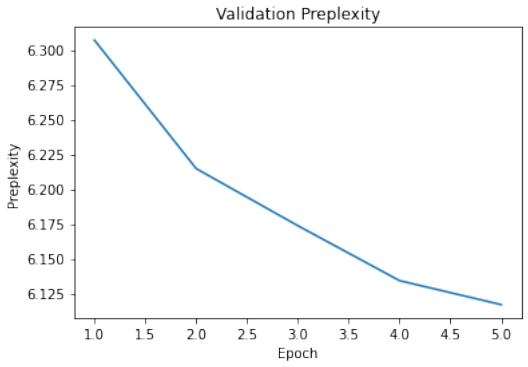

Figure 4: Perplexity Curve of GPT-2 (Few-Shot).

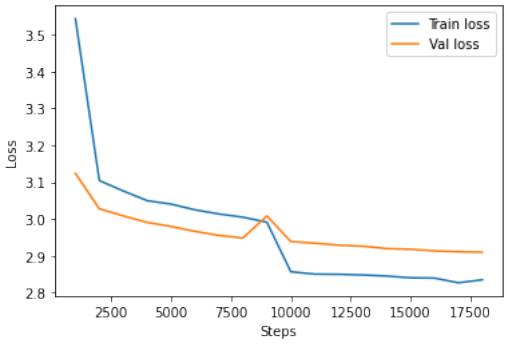

Figure 5: Loss Curves of BertSumEXT-BART Training.

Source	Demo
Article 1	(CNN)President Barack Obama took part in a roundtable discussion this week on climate change, refocusing on the issue from a public health vantage point. After the event at Washington's Howard University on Tuesday, Obama sat down with me for a one-on-one interview. I asked him about the science behind climate change and public health and the message he wants the average American to take away, as well as how enforceable his action plan is. ... He credits the Clean Air Act with making Americans "a lot" healthier, in addition to being able to "see the mountains in the background because they arenŕ covered in smog." ... In many ways, Obama is attempting to reframe the discussion around climate change as a public health issue that affects all of us, while conceding that we don't fully understand the magnitude of the correlation between rising temperatures and impact on human health. When asked what the average American can do about all this, the President encouraged ordinary citizens, doctors and nurses to start putting some pressure on elected officials "to try and make something happen to reduce the impacts of climate change." ...
GPT-2 (few-shot)	Obama is reframing the debate around climate change as a public health issue . He wants ordinary citizens to start putting pressure on elected officials to try and make something happen . He also urges ordinary citizens to start putting pressure on elected officials to try and make something happen . The President also urges ordinary citizens to start putting pressure on elected officials to try and make something happen . The President also urges ordinary citizens to start putting pressure on elected officials to try and make something happen . The President also urges ordinary citizens to
BART (ours)	President Obama spoke at a roundtable discussion on climate change this week. He said the Clean Air Act has made Americans "a lot" healthier. The average American can also do their part to reduce their own carbon footprint, he said. Obama did not appear particularly concerned about the Supreme Court challenge to the Affordable Care Act.
BERTSumEXT + GPT-2	President Obama says he wants the average American to take away from climate change. Obama says he's been "a lot" healthier than before. Obama says the Clean Air Act is a "public health issue" Obama says the EPA has been "a lot" healthier than before . Obama says the EPA is "a lot" better at regulating pollution than the EPA . Obama says the EPA has "a lot" better at regulating pollution than the EPA .Obama says the Clean Air
BERTSumEXT + BART	President Barack Obama sat down with CNN's John Sutter for a one-on-one interview. Sutter asked him about the science behind climate change and public health and the message he wants the average American to take away. Obama: "No challenge poses more of a public threat than climate change" He credits the Clean Air Act with making Americans healthier.
Article 2	Arsenal target Raheem Sterling is a more talented player than Gunners star Mesut Ozil, according to Michael Owen. The former Liverpool forward made the claim before Brendan Rodgers' men lost 4-1 at the Emirates on Saturday - **a game in which Ozil netted his side's second goal.** Owen's opinion drew criticism on social media but the ex-England man has refused to back down, even after Sterling's disappointing display in north London. Raheem Sterling, pictured against Arsenal, is a better player than Mesut Ozil according to Michael Owen . Ozil bends home Arsenal's second goal in their 4-1 win against Liverpool at the Emirates on Saturday. ... **Ozil celebrates after netting only his fourth Premier League goal of the season against Liverpool.**
GPT-2 (few-shot)	Arsenal target Raheem Sterling is a more capable player than Mesut Ozil . Raheem Sterling has scored twice in the Premier League this season . Raheem Sterling has scored twice in the Premier League this season . Raheem Sterling has scored twice in the Premier League this season . Raheem Sterling has scored twice in the Premier League this season . Raheem Sterling has scored twice in the Premier League this season . Raheem Sterling has scored twice in the Premier League this season
BART (ours)	Michael Owen believes Raheem Sterling is a better player than Mesut Ozil. Ozil scored Arsenal's second goal in their 4-1 win against Liverpool on Saturday. Sterling scored just his fourth Premier League goal of the season against Arsenal. Click here for all the latest Liverpool news.
BERTSumEXT + GPT-2	Raheem Sterling makes the claim after Arsenal beat Liverpool 4-1 at the Emirates. Arsenal's Raheem Sterling is a more talented player than Mesut Ozil, according to Owen. Towards the end of the game, Sterling scores his second goal. Arsenal's Raheem Sterling is a more talented player than Mesut Ozil, according to Owen. Towards the end of the game, Sterling scores his second goal. Towards the end
BERTSumEXT + BART	Michael Owen believes Raheem Sterling is more talented than Mesut Ozil. The former Liverpool forward made the claim before Arsenal's 4-1 win. Owen's opinion drew criticism on social media but the ex-England man has refused to back down, even after Sterling's disappointing display in north London.

Table 3: Example summaries from our fine-tuned models. Factually inconsistent sentences are highlighted in red.

Source	Demo
Article 3	(CNN)In her 40 years living in Rochelle, Illinois, Cathy Olson had never seen a tornado that big. "I saw the top of the funnel cloud, and it was absolutely massive," she said. She watched the hulking gray twister grind past her town Thursday, tearing up its fringes. Farther north, in the rural Illinois hamlet of Fairdale, one person died as a twister shredded homes and ripped trees bare of leaves and most limbs. Only the thickest branches remained standing. It was the only death reported so far in two days of tornado touchdowns. ... Multiple tornadoes ripped through the rural Midwest on Thursday. A large and dangerous twister tore across fields in Iowa. And a twister touched down 70 miles outside of St. Louis. Eight tornadoes were reported Wednesday in Oklahoma, Kansas and Missouri, the Storm Prediction Center said. But it appears residents have been fortunate enough to come away from the terrifying weather spectacle alive. Hail stones the size of tennis balls plummeted down on Ashton, Illinois. It could have been worse as severe tornado damage dotted a path not far from the dense populations of Chicago and Rockford – the stateś third largest city. The tornado cut a 22-mile path through Ogle County, according to disaster management coordinator Tom Richter. ...
GPT-2 (few-shot)	The hulking gray twister churned past her town Thursday . It was the only death reported so far in two days of tornado touchdowns . It was the only death reported so far in two days of tornado touchdowns . It could have been worse as severe tornado damage dotted a path not far from the dense populations of Chicago and Rockford . Hail stones the size of tennis balls fell down on Ashton , Illinois . It could have been worse as severe tornado damage dotted a path not far from the dense populations
BART (ours)	One person dies in Fairdale, Illinois, as a twister shreds homes and trees. "I saw the top of the funnel cloud, and it was absolutely massive," a witness says. The twister cut a 22-mile path through Ogle County, Illinois. It was the first tornado the sheriff had ever seen in his county.
BERTSumEXT + GPT-2	A tornado ripped through Iowa, Kansas and Missouri. Eight tornadoes were reported in Oklahoma, Kansas and Missouri. The tornado touched down 70 miles outside of St. Louis. The storm could have been worse as severe tornado damage dotted a path not far from the dense populations of Chicago and Rockford . The storm could have been worse as severe tornado damage dotted a path not far from the dense populations of Chicago and Rockford . A tornado touched down 70 miles outside of St. Louis
BERTSumEXT + BART	Multiple tornadoes rip through the rural Midwest on Thursday. A large and dangerous twister tore across fields in Iowa. Hail stones the size of tennis balls plummeted down on Ashton, Illinois. One person died as a twister shredded homes and ripped trees bare of leaves.
Article 4	The Sigma Alpha Epsilon (SAE) fraternity at Clemson University in South Carolina has been put on probation for two years after the group held a Christmas theme party that flared up racial tensions on campus. The 'Cripmas' party, held last December, had white students throwing stereotypical gang symbols while dressed in red and blue bandanas, T-shirts with images of the late rapper Tupac Shakur and fake 'thug' tattoos. The party was also reportedly attended by female students from several sororities. Photos from the party flooded social media and were accompanied by comments such as: 'Merry Cripmas to all, and all a hood night.' The Sigma Alpha Epsilon (SAE) fraternity at Clemson University in South Carolina has been put on probation for two years after the group held a Christmas theme party that flared up racial tensions on campus . Clemson University president, James Clements (right), speaks with students protesting last December's frat party which caused outrage for its perpetuation of stereotypes . The party, organized by several SAE members, was held off-campus at a private residence. 'Cripmas' appeared to be a reference to a primarily African-American gang, The Crips, founded in California. The incident caused a backlash as black students protested and said Clemson didn't do enough to promote racial tolerance. About 6 percent of Clemson's students are black. All SAE activity was suspended by Clemson indefinitely following the inappropriate theme party. ...
GPT-2 (few-shot)	The Sigma Alpha Epsilon 's fraternity at Clemson University has been put on probation for two years . The party was held last December . The party was attended by female students from several sororities . The party was also attended by female students from several sororities . The party was also attended by female students from several sororities . The party was also attended by female students from several sororities . The party was also attended by female students from several sororities . The party was also
BART (ours)	The Sigma Alpha Epsilon (SAE) fraternity at Clemson University in South Carolina has been put on probation for two years. The group held a Christmas theme party that flared up racial tensions on campus. White students dressed in red and blue bandanas, T-shirts with images of the late rapper Tupac Shakur and fake 'thug' tattoos. Photos from the party flooded social media and were accompanied by comments such as: 'Merry Cripmas to all'
BERTSumEXT + GPT-2	Sigma Alpha Epsilon fraternity held Christmas party in South Carolina. Party was attended by female students from several sororities. Party sparked racial tensions on campus. Party was held off-campus at a private residence. Club will be on probation until February 2017. Club will be on probation until February 2017.
BERTSumEXT + BART	Sigma Alpha Epsilon (SAE) fraternity at Clemson University in South Carolina has been put on probation for two years after the group held a Christmas theme party. The 'Cripmas' party, held last December, had white students throwing stereotypical gang symbols while dressed in red and blue bandanas, T-shirts with images of the late rapper Tupac Shakur and fake 'thug' tattoos. An internal investigation found that the fraternity violated alcohol rules and student conduct codes.

Table 4: Example summaries from our fine-tuned models. The articles are abbreviated.