Identifying Factors Impacting News Summarization Performance of Pre-trained Transformer Models

Ali Nikooyan
anikooyan@berkeley.edu

Julia Ying
julia_ying@berkeley.edu

Abstract

News industry has a growing need for automatically generated high quality news summaries. Although advanced pre-trained transformer models have achieved decent overall performance on abstractive news summarization, performances on individual articles vary greatly. This study aims to investigate the article-specific factors that have the most impact on the quality of model generated summaries. Results suggest the writing styles of news articles have a considerable correlation with the level of similarity between model generated summaries and references. Having key information in earlier parts of the article tends to increase abstractive summarization performance. Consequently, input news articles for summarization models can be truncated significantly without notable loss of performance as measured by ROUGE.

1 Introduction

With the emergence of technology and consequently increasingly fast-paced life-style, people are exhibiting shorter attention spans. As a result, there has been a decrease in demand for reading, followed by decreased value for news markets (Singh et al., 2021). Therefore, the ability to summarize news articles into succinct and informative snippets would enable a much more effective way of delivering content through media appropriate for modern day information consumption.

Over time, summarization has been widely adopted into various industries and there has been a growing transition from extractive towards abstractive summarization. While in the former, the most informative sentences are ranked and entirely copied from the original text, the latter can generate novel words and sentences that not only can cover the critical concepts of the entire text, but also linguistically sound.

Despite recent breakthroughs in summarization model complexity and performance, little has been done on evaluating how the composition of the sources or topics of the training text and/or the characteristics of articles used for either training or summary generation can affect the performance of these models. These topics are equally important as the complexity of the models, since the training curriculum could be critical to improving summary generation on a large scale, in particular for specialized applications across a wide range of industries.

2 Background

Since its introduction in 2014 (Sutskever et al., 2014), the sequence-to-sequence architecture using encoder-decoder based on either RNNs or transformers has become the dominant architecture for abstractive summarization (Chung et al., 2014; Vaswani et al., 2017). Several transformer-based sequence-to-sequence models have been developed and adapted for text generation in general and specifically abstractive summarization.

Summarization models requires large text datasets for pre-training and fine-tuning. Examples of large corpora from major publications include HugeNews (1.5B articles collected from news websites), XSum (227k BBC articles)(Narayan et al., 2018), CNN/DM (93k articles from CNN, and 220k articles from DailyMail)(Hermann et al., 2015), the New York Times Annotated Corpus (1.8M articles with 650k summaries by NYT)(Sandhaus, 2008), NEWSROOM (1.3M article from 38 major sources)(Grusky et al., 2018), and Multi-News (56k pairs of news article from newser.com)(Fabbri et al., 2019).

Majority of recent studies on text summarization have focused on improving the performance, as measured by the ROUGE score (Lin, 2004), by enhancing the architecture of the models, developing hybrid fine-tuning strategies, trying novel masking techniques, and/or increasing the size and diversity of the datasets used for pre-training.

BERTSumExtAbs (Liu and Lapata, 2019), an advanced model built on top of BERT (Devlin et al., 2019) specializing in summarization, was able to achieve ROUGE-1 F1 of 42.13 tested on CNN/DM dataset by implementing a two-stage encoder fine-tuning approach, first with an extractive objective and subsequently abstractive summarization.

Table 1: Previously reported ROUGE Scores in literature for different pre-trained models

Model architecture	Pre-trained on	Fine-tuned on CNN/DM			Fine-tuned on NEWSROOM		
		R1 - F1	R2 - F1	RL- F1	R1 - F1	R2 - F1	RL - F1
$\textbf{T5}_{\text{LARGE}}$	C4	43.52	21.55	40.69	-	-	-
$\textbf{PEGASUS}_{\text{LARGE}}$	C4	43.90	21.20	40.76	45.07	33.39	41.28
	HugeNews	44.17	21.4	41.11	45.15	33.51	41.33
$\textbf{BART}_{\text{LARGE}}$	5 Corpora (Liu et al., 2019)	44.16	21.28	40.90	-	-	-

By scaling up model size to 11 billion parameters and a massive text corpus derived from Common Crawl (C4), T5 model (Chung et al., 2014) showed slight improvement in R1 (43.33, Table 1) after fine-tuning on CNN/DM.

The Bidirectional and Auto-Regressive Transformers (BART) (Lewis et al., 2020), a denoising autoencoder built with a sequence-to-sequence model that corrupt text with an arbitrary noising function and then learns to reconstruct the original text, improved R1 to 44.16 by fine-tuning on CNN/DM. (Table 1).

The Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) model (Zhang et al., 2019) has achieved the current state-of-the-art performance (R1 = 44.17 with pre-trained model on CNN/DM, Table 1). The major advancement in PEGASUS lies in masking whole sentences from a document instead of smaller continuous text spans as its peers do. Both BART and PEGASUS show better performance in abstractive summarization when fine-tuned on news datasets with more abstractive summaries comparing to when fine-tuning with more extractive summaries, suggesting importance of training curriculum.

3 Experiment Setup

3.1 Research Questions

In this study, we aim to explore how characteristics of the training curriculum impact summarization models' performance, focusing on answering the following questions:

1. By using an improved training curriculum, can we outperform the state-of-the-art model in producing high quality summaries?
2. How much would truncating input articles impact summary generation?
3. What characteristics of training and testing texts impact summarization performance?

3.2 Dataset

The CORNELL NEWSROOM dataset (Grusky et al., 2018) is an open source dataset including 1.3 million articles by 38 major publications between 1998 and 2017. Corresponding summaries were obtained from search and social metadata, and consists a variety of summarization strategies utilizing both abstractive and extractive summarization. The diverse characteristics within this dataset makes it an ideal candidate for us to explore aforementioned research questions. In addition to the fields available directly from the raw data, we also parsed the news outlet and news categories through the original news url link provided in the raw data. Note only 16 of the 38 news sources in the dataset indicate the article's category in its url.

The goal of this project is to train a model to produce high quality summaries, which we define as a succinct message of approximately 1 to 3 sentences that captures important information from the original text. Furthermore, as news articles tend to adhere to the *inverted-pyramid* writing style where core ideas are concentrated in the beginning of the article, we strive for an abstractive model that does not exploit the *head-heavy* writing style and simply recite the beginning passage of the article verbatim.

Past studies on summarization using the NEWSROOM dataset (Zhang et al., 2019) did not apply data cleaning to remove poor examples of summaries. For our purpose of training the model to produce high quality abstractive summaries, we employed aggressive data cleaning and removed the following types of records:

- missing the news article or summary
- duplicate summaries or articles
- in a language other than English
- contains a hyperlink
- summaries shorter than 5 or longer than 135 words (2nd and 98th percentile, respectively)
- articles shorter than 100 or longer than 1000 words
- reference summaries with ROUGE-2 recall > 0.15 (subsequently referred to as *RefSum ROUGE*) when evaluated against the first 150 words of the article, as vast majority of these

summaries were copying the beginning of the article

- articles and summaries with { or } which indicates a bad parse of the web crawl
- articles with more than 3 repeating 5-grams, mostly photo gallery posts with repeating identical captions.

Appendix A shows examples of records removed during data cleaning. The 592,750 articles remaining were randomly split into 400,000 train, 100,000 development, and 92,750 test sets.

3.3 Model

BART is a seq2seq model with a bidirectional encoder and a left-to-right decoder. Model pre-training includes the corruption of text with an arbitrary noising function, and then learning to reconstruct the original text. BART's key feature, noise flexibility, makes it suitable for abstractive summarization on a noisy dataset like NEWSROOM. We opted to use one of the pre-trained BART model variants available on Hugging Face (Wolf et al., 2020), which consists of 12 encoder and 12 decoder layers. The pre-trained model, designated as *bart-large-cnn*, had been fine-tuned on CNN/DM.

To establish a comparable baseline performance, we also utilized the *pegasus-large* model fine-tuned on NEWSROOM and pre-trained *t5-base*.

3.4 Methodology

We used an AWS p4d.24xlarge instance with 8 NVIDIA A100 GPU to fine-tune BART on the cleaned NEWSROOM dataset. For tokenization, we used the default tokenizer associated with each model from Hugging Face. During tokenization, the articles were truncated or padded to the maximum input length allowed by the model (1024 tokens). To reduce computation demands, weights in the 12 encoder layers were frozen. For the basic fine-tuning, we trained the model for 1.25 epochs in batch size of 10 per GPU, with 5 validation cycles in each epoch. The full training/validiation process took approximately 5 hours.

We further optimized the hyperparameters using the built-in hyperparameter search method of Hugging Face's Trainer class with Ray library as the backend. Working with 25% of the training data, we searched through 10 random combinations of learning rate, weight decay rate, and random seed, and chose the one yielding the lowest validation loss. The hyperparamter search process lasted over 15 hours. We then repeated the fine-tuning process using the optimal hyperparameter found. Based on training and validation loss, the training lasted two epochs, approximately 8 hours.

For test article summary generation using the fine-tuned BART model, T5, and PEGASUS, the inputs were tokenized to the maximum length each model allows (512 for PEGASUS and T5, 1024 for BART).

To evaluate effects of input article length, we truncated the tokenized articles in two ways. In forward truncation, a percentage of non-padding tokens at the end in each input is replaced with padding tokens, and the overall input length is trimmed proportionally. In backward truncation, a percentage of non-padding tokens at the beginning of each input is replaced with padding tokens, and the first of remaining non-padding token is replaced with the beginning-of-sentence token.

Inputs for BART were forward truncated by 50%, 75%, or 87.5%, and oveall input lengths were trimmed to 512, 256, 128 tokens respectively. Inputs for PEGASUS were forward truncated by 50% or 75%, and overall input lengths were trimmed to 256 and 128 tokens respectively. For backward truncation, inputs for BART were truncated by 12.5%, 25%, or 50%, while inputs for PEGASUS were truncated by 25% or 50%.

3.5 Metric

To measure the similarity between the model generated summaries and reference summaries, we used ROUGE-1 (R1) and ROUGE-2 (R2), as well as ROUGE-L (RL). While recall is the main focus of the ROUGE statistic, other forms of ROUGE scores have been introduced by considering the length of the generated summary (Precision) or both recall and precision (F1). There is no consensus in the literature about which ROUGE scores to report. While the studies we cited here prioritized ROUGE F1, we report all three forms (recall, precision, and F1) of each ROUGE score, but prioritize recall over F1 when it comes to conclusions about model performance. This choice is motivated by our primary goal of preserving maximum information on the reference summary side.

4 Results and Discussions

4.1 Effects of Training Curriculum

ROUGE recall and F1 scores obtained on the test set are summarized in Table 2.

Table 2: ROUGE scores (R1/R2/RL) for summaries generated by various models. Maximum length considered in ROUGE score evaluation is 150 words. Summary generation used beam size ranging from 2-5, though only the beam size yielding the best RL is reported.

	Beam size	Recall	F1
T5	4	41.95 / 15.78 / 37.92	20.48 / 7.63 / 21.01
PEGASUS	5	**52.79 / 28.45 / 48.81**	**30.28 / 16.52 / 30.88**
BART without fine-tune	3	43.83 / 16.47 / 39.05	23.49 / 8.68 / 23.28
BART fine-tuned	2	50.50 / 24.97 / 45.96	28.63 / 14.10 / 28.71
BART fine-tuned with optimal HP	2	51.67 / 25.72 / 46.81	29.13 / 14.50 / 29.14

Performance of pre-trained BART without fine-tuning was similar to T5, the lowest across all three ROUGE recall scores. Fine-tuning on Newsroom considerably boosted the performance of BART-base model. By fine-tuning BART without using optimal hyperparameters, R1 / R2 / RL increased by 6.67, 8.5, 6.91 points, respectively. By using the optimal fine-tuning, we were able to further increase ROUGE scores by about 1 point each.

Though ROUGE score is a common metric for summarization, published literature do not typically disclose the conditions under which the scores were obtained, such as beam size or maximum length to consider in ROUGE calculation. Moreover, because our dataset has been significantly pruned and therefore have different characteristics from test sets used by other studies, our ROUGE scores are not directly comparable to that of prior studies. For this reason, the more suitable baseline is the ROUGE scores of summaries generated by the state-of-art model, PEGASUS, using the same test set and under the same condition.

Notably, our PEGASUS baseline has significantly lower ROUGE F1 scores compared to prior studies (Table 2). Prior reported numbers may be conflated by including summaries directly quoting the early portion of the article, which are mostly left out from our data after data cleaning.

Our optimal fine-tuned model underperformed the PEGASUS baseline by a only slim margin (1.12 R1, 2.73 R2, 2.00 RL). However, analysis of BART and PEGASUS generated summaries shows a different distribution of summary length (Fig. 1). Token counts of BART summaries display a unimodal distribution and conform to our intended summary length (min = 25, max = 140, mean = 55.11, sd = 8.11). In contrast, PEGASUS summaries have a bimodal distribution with a long tail (min = 5, max = 428, mean = 64.65, sd = 46.82). The longer summaries by PEGASUS might contribute to the higher recall scores, and the cluster of short summaries contributes to higher F1/precision scores.

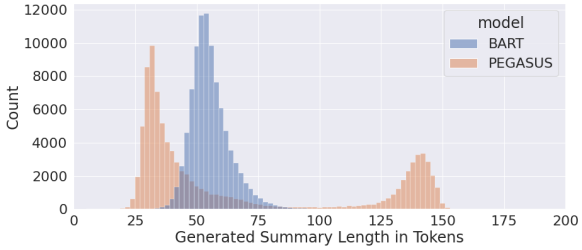


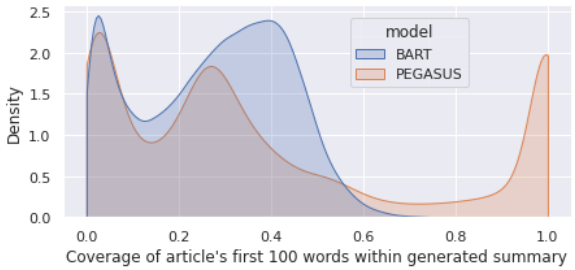
Figure 1: Histogram of generated summary length

Next, we assessed how much the models rely on the earlier sections of input articles when generating summaries. The first 100 words of the articles were used as reference to evaluate R2 recall of the generated summaries. Conceptually, if R2 is high, more texts in the beginning of article are covered by the generated summary, and by extension, the summary is mainly gathered from the beginning of the article.

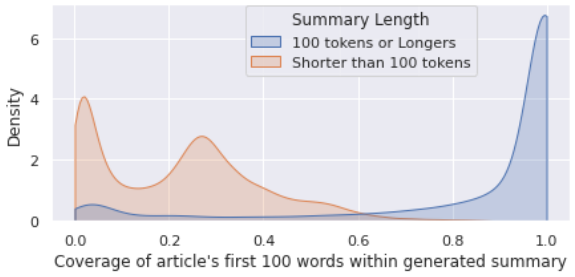
Density estimation (Fig. 2a) shows PEGASUS having a large peak at R2 = 1. Subdividing PEGASUS summaries based on length suggests that the longer summaries are essentially copying the entire beginning section of the article as majority have R2 close to 1 (Fig. 2b). Subdividing BART summaries at the median length does not show significant R2 disparity between the two subgroups.

4.2 Effects of Reducing Input Length

To evaluate the effects of reducing input article length on summary generation, we applied the truncation treatment to test articles as described in Section 3.4 during summary generation using both PEGASUS and fine-tuned BART. Both models show similar patterns, with small reduction in performance when trimming the input articles up to 50% from the end (forward truncation). Performance experiences a sharp drop (up to 5 R2 recall scores) if truncated further. When truncating from the beginning (backward truncation), performance drops considerably (up to 15 R2 recall) for the first



(a)



(b)

Figure 2: (a) Coverage of beginning sections of articles by generated summaries. (b) Coverage by PEGASUS generated summaries grouped by length

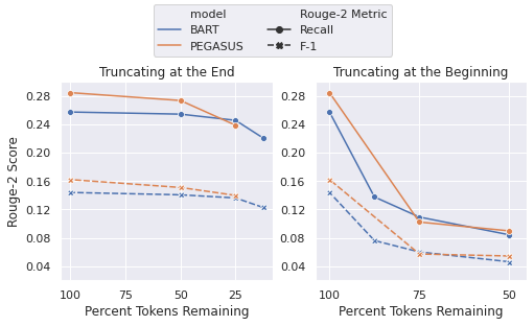


Figure 3: R2 score for various truncation conditions. R1 and RL showed highly similar patterns

25% and suffers smaller drops in subsequent truncation, further highlighting the emphasis of article's opening paragraphs during summarization. (Table 3, Fig. 3). Our observations are consistent with a previous study on PEGASUS comparing three different methods of gap-sentence masking during pre-training (Zhang et al., 2019). Masking articles' lead sentences led to lower performance comparing to the other two masking strategies.

To investigate possible connections between performance loss during truncation and length of the article, we calculated the delta of mean R2 change from no truncation to 75% forward truncation. Correlation of delta and article length or reference summary length were negligible. Scatter plot of 1000 random samples ruled out any obvious non-linear relationship. This would imply the drop in performance due to forward truncation is uniform across different article lengths.

However, the delta of mean R2 change from no truncation to 25% backward truncation showed a moderate correlation to the article length (-0.1 for both BART and PEGASUS). The negative correlation suggests that when backward truncating input article by 25%, the drop in performance is smaller for longer articles and greater for shorter articles. It is likely that shorter articles adhere more to the inverted pyramid writing style where the important information is concentrated in the beginning (head-heavy), whereas longer articles might include more expositions and have key information more spread out in the article.

In summary, news articles are for most part, "head-heavy". The key findings here indicate it would be feasible to truncate the input for BART to 512 tokens (half of maximum allowed) for efficiency without significant loss of ROUGE score.

Table 3: ROUGE scores (R1/R2/RL) of summaries generated under various forward (F) and backward (B) truncation

	Max Tokens	Beam size	Recall	F1
PEGASUS - no trunc.	512	5	52.79 / 28.45 / 48.81	29.78 / 16.20 / 30.40
PEGASUS - 50% F trunc.	256	5	51.95 / 27.34 / 47.88	28.45 / 15.09 / 29.12
PEGASUS - 75% F trunc.	128	5	46.89 / 23.86 / 43.64	27.31 / 13.97 / 27.94
PEGASUS - 25% B trunc.	1024	5	38.44 / 12.64 / 35.29	22.98 / 7.81 / 22.98
PEGASUS - 50% B trunc.	1024	5	33.71 / 8.97 / 31.38	19.56 / 5.44 / 19.90
BART - no trunc.	1024	2	51.67 / 25.72 / 46.81	29.03 / 14.39 / 29.08
BART - 50% F trunc.	512	2	51.14 / 25.42 / 46.34	28.44 / 14.05 / 28.55
BART - 75% F trunc.	256	2	49.56 / 24.55 / 45.22	27.69 / 13.61 / 27.95
BART - 87.5 F trunc.	128	2	45.59 / 22.06 / 42.44	25.64 / 12.28 / 26.35
BART - 12.5% B trunc.	1024	2	41.67 / 13.75 / 37.03	23.19 / 7.61 / 22.74
BART - 25% B trunc.	1024	2	38.72 / 10.92 / 34.49	21.47 / 6.01 / 21.10
BART - 50% B trunc.	1024	2	35.28 / 8.42 / 31.81	19.56 / 4.62 / 19.44

4.3 Effect of News Source and Article Type

To investigate the impacts of new articles' characteristics, we considered potential factors influencing the model performance (as measured by R2 recall) including head-heaviness, reference summary length, article length, and article to reference summary length ratio. RefSum ROUGE, as defined in Section 3.2 is used as a proxy for head-heaviness, i.e. the extent to which the reference summary would co-occur with the first 150 words of the article.

Correlation analysis using generated summaries from test articles shows the head-heaviness as the only factor to be highly and positively correlated with the model performance (Fig. 4a). Next, we contrasted records where BART performed well versus poorly. The high and low performance groups consist of records whose generated summary R2 is in Q4 and Q1, respectively. Density estimation plot of the two groups (Fig. 4b) shows that the majority of low-performing summaries have very low head-heaviness, suggesting the writing style of an article is a major predictor of the quality of BART generated summary.

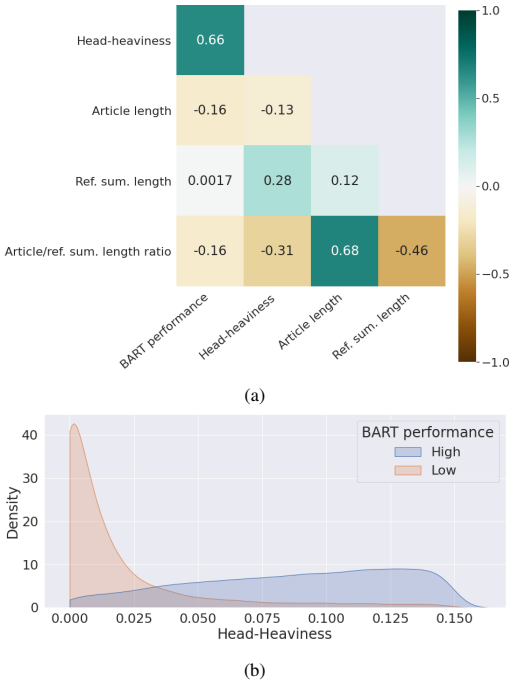


Figure 4: (a) Correlation matrix of potential factors influencing model performance. (b) Density estimation of writing style head-heaviness by high and low model performance

To further examine the association between head-heaviness and the model performance, we performed a correlation analysis for each news source/category in our test dataset. Only news sources with more than 1000 records and article categories with more than 500 records in the test set were considered in order to avoid small samples that are not representative. The analysis includes 21 news sources and 24 different categories.

For news source/article category, R2 recall, as well as RefSum ROUGE were calculated and averaged across all articles in that source or category. Normalized R2 and RefSum ROUGE are visualized in Fig. 5, which shows that sources or categories with highest R2 tend to have highest RefSum ROUGE scores; conversely, sources or categories with lowest R2 tend to have lowest RefSum ROUGE scores. In other words, the BART generated summaries are closer to their human generated reference for news sources or article categories that are more head-heavy. On the other end of the spectrum, categories such as opinions and comments have some of the lowest mean R2, and these types of articles tend to be expositions that do not adhere to the inverted-pyramid writing style typical of news articles. This finding draws attention to the potential need for multiple models to specialize in summarizing different writing styles instead of attempting to accommodate all writing styles using a generalized model.

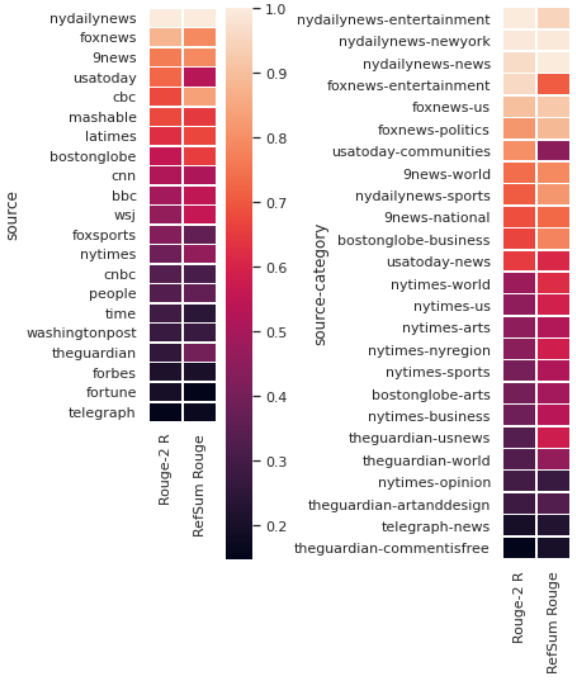


Figure 5: Heatmap of normalized mean R2 recall for BART generated summaries evaluated against reference summary, and RefSum ROUGE, a proxy for "head-heaviness" of articles, grouped by news outlet (left) and news category (right).

5 Conclusions

After fine-tuning on cleaned NEWSROOM dataset, the pre-trained BART-large model achieved performance comparable to the state-of-the-art model in generating high quality summaries. Moreover, truncating input articles from the end by up to 50% was found to have negligible effect on the quality of the generated summaries. This is an important finding that can help saving time and resources for the downstream summarization tasks in the future. Finally, the results in this study showed that the inverse-pyramid writing style is critical in achieving good abstractive news summarization with the BART model.

5.1 Challenges and Limitations

Computing resource scarcity and time limitation were the major limiting factors for fine-tuning the pre-trained transformer model. If allowed more time and resources, we would perform a more rigorous hyperparameter-tuning, include the model's encoder layers in the training process, and experiment with different model architectural enhancements for the downstream specialized summarization task.

Another major challenge universal to the summarization task is that all forms of ROUGE score may not be an accurate measure of how informative a generated summary would be about the underlying text, as we have observed model generated summaries that look more informative than their human generated counterparts. Appendix B shows some examples of meaningful generated summaries yielding low R2, and summaries with high R2 that are merely copying the beginning of the article and lacking depth of information.

Despite achieving good performance on news summarization tasks, our fine-tuned model will be limited in scope and generalizability. It would not perform well on texts not written in the inverted-pyramid style. Appendix C shows examples of summaries of classic literature generated by both PEGASUS and BART. PEGASUS, due to the nature of the whole NEWSROOM dataset, defaults to quoting the first few lines. But because BART was fine-tuned on data that discourages summaries emphasizing the opening sentences, it instead "plucks" words scattered across the input text and assembles a summary that although grammatically sound, completely diverges from the semantic content of original text.

5.2 Future works

Based on findings in this study, the future researches should be directed towards specialized abstractive summarization of news articles weighing-in the characteristics of the article such as the writing style, the news source, and the article category as well as the development of more representative metrics other than ROUGE for assessing the quality of the model generated summaries.

References

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Rajeev Kumar Singh, Sonia Khetarpaul, Rohan Gorantla, and Sai Giridhar Rao Allada. 2021. Sheg: summarization and headline generation of news articles using deep learning. *Neural Computing and Applications*, 33.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *the 31st International Conference on Neural Information Processing Systems*, page 6000–6010. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

A Examples of records removed during data cleaning

- Summaries or articles in foreign language

  Language was determined using the `langid` library on first 150 words of the article.

  "Kwamitin kula da wasannin Olympics na duniya, IOC, zai yanke kodai za a hana Russia shiga gasar wasannin Olympics ta Rio ko a a."

- Summaries containing urls

  "The banning of `&#8220; <a href="http://www.youtube.com/watch?feature=playerembedded &#038;v= udUBKu4go7s"> Fuelling Poverty</a> &#8221;` leaves many Nigerians questioning the government of Goodluck Jonathan."

- Summaries shorter than 6 words

  "Florida Gov"
  "Come see"

- Summaries copying article verbatim

  "If you're cutting back on stocks because interest rates are rising, you're making a mistake. But don't just take my word for it (after all, I am a dyed-in-the-wool dividend-stock fan). Ask Ned Davis Research, which released its latest research on the relationship between stocks and rates about a year ago. The [...]"

- Badly parsed article or summaries

  " `{"asset_collection": [{"headline": "McAfee repeats innocence claim", "photo": {"crops": {"1_1": "http://www.gannett-cdn.com/media/USATODAY/USATODAY/2012/12/14/afp-515993850-1_1.jpg"}}` "

- Articles or summaries with too many repeating n-grams

  "Oscars 2015: Red carpet 50 photos

  Zendaya arrives on the red carpet for the 87th Academy Awards on Sunday, February 22.

  Oscars 2015: Red carpet 50 photos

John Travolta and Scarlett Johansson

Oscars 2015: Red carpet 50 photos

Jennifer Aniston, left, and Emma Stone

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

Lady Gaga, left, and Keira Knightley

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos"

- Other unmeaningful articles or summaries

"Browse stories from Carey Reilly on FoxNews.com."

"ARF ARF ARF ARF ARF GRRRR. ARF ARF."

B Examples of BART generated news summaries

B.1 Example 1

ROUGE-2 Recall: 0.0

generated summary:

Two men left stunned after being asked to follow security guard out of Sainsbury's after he complained about their holding hands and putting their arms around each other in the aisles of a London branch of the supermarket chain. The pair were offered an apology and a £10 voucher for their troubles

reference summary:

Supermarket says sorry but gay couple urge better training for staff after their treatment at branch in Hackney, east London

article exerpt:

Sainsbury's has apologised to two men who say they were asked to follow a security guard out of a shop after another customer complained about them holding hands and putting their arms around each other.

Thomas Rees said he and his partner, Joshua Bradwell, had been left stunned after the security guard led them out of a Sainsbury's branch in Hackney, east London, and told them a woman had complained about their behaviour in the aisles.

After he tweeted his anger at the supermarket chain, he was offered an apology and a £10 voucher. But he has said he now feels uncomfortable at the prospect of returning to the shop and running into the security guard again or discovering other customers with intolerant views.

Sainsbury's has apologised to two men who say they were asked to follow a security guard out of a shop after another customer complained about them holding hands and putting their arms around each other.

Thomas Rees said he and his partner, Joshua Bradwell, had been left stunned after the security guard led them out of a Sainsbury's branch in Hackney, east London, and told them a woman had complained about their behaviour in the aisles.

After he tweeted his anger at the supermarket chain, he was offered an apology and a £10 voucher. But he has said he now feels uncomfortable at the prospect of returning to the shop and running into the security guard again or discovering other customers with intolerant views.

B.2 Example 2

ROUGE-2 Recall: 0.0

generated summary:

China's consumer price index rose 5.1 percent in November compared with the same month in 2009, the sharpest increase in three years, and job satisfaction was at its lowest in four years, according to a survey of 4,143 people by the Chinese Academy of Social Sciences.

reference summary:

The comments from Li Keqiang, vice premier of China, seemed to highlight that China is increasingly concerned about resentment over rising living costs.

article exerpt:

The government reported on Dec. 11 that the consumer price index rose 5.1 percent in November compared with the same month in 2009; it was the sharpest increase in three years. Chinese planners had hoped that the average of the index for the year would not go above 3 percent, but that looks virtually impossible now. Consumers have been especially hard hit by soaring food prices.

On Wednesday, a prominent research organization, the Chinese Academy of Social Sciences, reported that high inflation and soaring housing prices had contributed to a growing sense of popular disaffection. The findings of the report, based

on a survey of 4,143 people, appeared in this year's edition of the Blue Book of China's Society.

Commodity prices were the main concern of urban residents, followed by health care and housing prices, according to the findings, which were reported by Xinhua on Thursday. Last year, the rise in commodity prices was listed as fifth among the concerns of urban residents. Rural residents in this year's survey said health care was their top concern, followed by commodity prices.

Job satisfaction among those surveyed was as its lowest in four years, according to the academy.

B.3 Example 3

ROUGE-2 Recall: 0.55

generated summary:

The veteran British actress, 81, took home the best supporting actress trophy for her role in the Kenneth Branagh Theatre Company's production of Paulina at London's Royal Opera House – her eighth Olivier Award win to her name. She is the most decorated actor in Olivier history

reference summary:

Judi Dench is now the most decorated actor in Olivier Awards history

article exerpt:

04/04/2016 AT 11:15 AM EDT

is an Olivier Award winner, yet again. And she notched a new record with her victory.

The veteran British actress, 81, took home the best supporting actress trophy during the

at London's Royal Opera House – her eighth win.

Dench is the most decorated actor in Olivier history, with the Sunday nod only solidifying her dominance. The Awards are presented once a year by the Society of London Theatre, and recognize excellence in the city's professional theatre. Audra McDonald is the only stage star to come near Dench's record, with six Tonys – the Olivier Awards' American equivalent – to her name.

B.4 Example 4

ROUGE-2 Recall: 1.0

generated summary:

Exhausted college students should consider their smartphone habits, according to a new study that links heavy texting with sleep problems and poor sleep habits, a new report suggests. Researchers from Washington and Lee University in Virginia found that students who send the most text messages had the poorest sleep habits and the lowest levels of emotional well-being.

reference summary:

Exhausted college students should consider their smartphone habits, according to a new study that links heavy texting with sleep problems.

article exerpt:

Exhausted college students should consider their smartphone habits, according to a new study that links heavy texting with sleep problems.

Researchers from Washington and Lee University in Virginia discovered that among first-year students, those who sent the most text messages had the poorest sleep habits and lowest levels of emotional well-being.

C Examples of generated literature summaries

C.1 Great Expectations by Charles Dickens

Original text excerpt: My father's family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip. So, I called myself Pip, and came to be called Pip.

I give Pirrip as my father's family name, on the authority of his tombstone and my sister,—Mrs. Joe Gargery, who married the blacksmith. As I never saw my father or my mother, and never saw any likeness of either of them (for their days were long before the days of photographs), my first fancies regarding what they were like were unreasonably derived from their tombstones. The shape of the letters on my father's, gave me an odd idea that he was a square, stout, dark man, with curly black hair. From the character and turn of the inscription, "Also Georgiana Wife of the Above," I drew a childish conclusion that my mother was freckled and sickly. To five little stone lozenges, each about a foot and a half long, which were arranged in a neat row beside their grave, and were sacred to the memory of five little brothers of mine,—who gave up trying to get a living, exceedingly early in that universal struggle,—I am indebted for a belief I religiously entertained that they had all been born on their backs with their hands in their trousers-pockets, and had never taken them out in this state of existence.

BART summary: The first time I saw my father and mother, I thought they were dead and buried in the churchyard of a church in the marshes of a bleak parish. I was a bundle of shivers growing

afraid of it all and beginning to cry, and began to call myself Pip.

PEGASUS summary: My father's family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip. So, I called myself Pip, and came to be called Pip

C.2 The Picture of Dorian Gray by Oscar Wilde

Original text excerpt: The studio was filled with the rich odour of roses, and when the light summer wind stirred amidst the trees of the garden, there came through the open door the heavy scent of the lilac, or the more delicate perfume of the pink-flowering thorn.

From the corner of the divan of Persian saddle-bags on which he was lying, smoking, as was his custom, innumerable cigarettes, Lord Henry Wotton could just catch the gleam of the honey-sweet and honey-coloured blossoms of a laburnum, whose tremulous branches seemed hardly able to bear the burden of a beauty so flamelike as theirs; and now and then the fantastic shadows of birds in flight flitted across the long tussore-silk curtains that were stretched in front of the huge window, producing a kind of momentary Japanese effect, and making him think of those pallid, jade-faced painters of Tokyo who, through the medium of an art that is necessarily immobile, seek to convey the sense of swiftness and motion. The sullen murmur of the bees shouldering their way through the long unmown grass, or circling with monotonous insistence round the dusty gilt horns of the straggling woodbine, seemed to make the stillness more oppressive. The dim roar of London was like the bourdon note of a distant organ.

BART summary: A portrait of a garden in London by Henry Wotton. The studio was filled with the rich odour of roses, and when the light summer wind stirred amidst the trees of the garden, there came through the open door the heavy scent of the lilac, or the more delicate perfume of the pink-flowering thorn.

PEGASUS summary: The studio was filled with the rich odour of roses, and when the light summer wind stirred amidst the trees of the garden, there came through the open door the heavy scent of the lilac, or the more delicate perfume of the pink-flowering thorn.