Domain Specific Fine-tuning of Denoising Sequence-to-Sequence Models for Natural Language Summarization

Exploring Dataset Augmentation with Domain-Specific Reddit Data and Optimal Fine-Tuning of BART for Text Summarization

Brydon Parker[1] Alik Sokolov[2] Mahtab Ahmed[3 5] Matt Kalebic[4]
Sedef Akinli Kocak[5] Ofer Shai[1]

[1] Deloitte [2] University of Toronto, Risklab [3]Western University [4]PwC Canada [5]Vector Institute

Abstract

Summarization of long-form text data is a problem especially pertinent in knowledge economy jobs such as medicine and finance, that require continuously remaining informed on a sophisticated and evolving body of knowledge. As such, isolating and summarizing key content automatically using Natural Language Processing (NLP) techniques holds the potential for extensive time savings in these industries. We explore applications of a state-of-the-art NLP model (BART), and explore strategies for tuning it to optimal performance using data augmentation and various fine-tuning strategies. We show that our end-to-end fine-tuning approach can result in a 5-6% absolute ROUGE-1 improvement over an out-of-the-box pre-trained BART summarizer when tested on domain specific data, and make available our end-to-end pipeline to achieve these results on finance, medical, or other user-specified domains.

1. Introduction

As the scale and variety of text data encountered in contemporary life continues to surge, effective summaries of information have become increasingly important. However, making the choice of what information to retain represents a substantial challenge that is both costly and time-consuming when performed manually, often making automation a requirement. Valuable text data often requires domain-specific knowledge to discern, which makes it difficult to crowd-source the parsing of such texts. This need for specialized readers or labelers can make the costs of natural language processing (NLP) projects on highly technical domains prohibitive. Motivated by recent advancements in NLP and model pre-training, our work seeks to advance automation of domain-specific text summarization through elucidating optimal end-to-end methodologies, from data collection to model fine-tuning.

Automated summarization is key opportunity in the field of NLP research for a number of industries, and especially in finance and healthcare. In finance, recent industry trends in investment research have emphasized this area in particular. Historically, highly specialized and experienced investment analysts have spent much of their time synthesizing structured reports from disparate news, financial statement and investor call data. Given advances in machine learning today coupled with regulatory changes and competitive forces, industry research teams are making a rapid push for automation in this area [27].

Similarly, healthcare is an area where a significant research burden is placed on highly qualified, educated, and experienced practitioners. As an example, the recent COVID-19 crisis alone has created the need to parse through a vast amount of scholarly articles. As of April 2019, roughly one month after the novel coronavirus was declared a pandemic by the World Health Organization (WHO), the Allen Institute for AI published a dataset of over 138,000 scholarly articles referencing COVID-19, SARS-CoV-2, and related coronaviruses [26]. When considering the need for healthcare practitioners to balance the necessity of remaining up-to-date against the sheer scale of information published, it is clear that NLP text parsing and summarization capabilities can be highly beneficial.

While extensive progress has been achieved in methods for performing extractive and abstractive summarization [1] [20] [10], reliable and sufficiently large datasets for model training and evaluation have remained relatively scarce. Efforts to date have largely focused on constructing large general annotated corpora based on news, such as DUC [6], Gigaword [14] [2], CNN / Daily Mail [13], and NEWSROOM [5], as well as from knowledge bases such as WikiHow [9] and longer-form documents such as research papers [3]. Recognizing the gap in specialized domain-specific datasets, our work leverages the unique community structure of Reddit, a prominent social news website that promotes discussion in forums (”subreddits”) devoted to specific topics. We capitalize on the use of ”TLDR” acronyms (”too long; didn’t read”), which indicate a concise summary of a lengthy comment. In this paper, we propose an improved extraction of Reddit TLDR’s relative to baseline methods [24]. We also provide tools for end-users to develop domain-specific text summarizers. We do this by sharing our framework for generating domain-specific training data by targeting specific subreddits, thereby facilitating fine tuning of pre-trained models and enhancing their overall performance

Lastly, we provide an analysis of abstractive summarization on datasets generated from finance and medical subreddits using Bidirectional and Auto-Regressive Transformers (”BART”) [10], a denoising autoencoder for pretraining sequence-to-sequence models that adopts and expands on Bidirectional Encoder Representations from Transformers (”BERT”) [4] and GPT pretraining [16] schemes.

2. Related Work

Approaches to text summarization can be classified into two main categories: *extractive* and *abstractive* [1]. Extractive summaries identify which sections of the text to retain versus discard, producing a cropped and stitched version of the original text with no alteration to phrasing. A classical approach to extractive summarization is using unsupervised methods such as Latent Semantic Analysis (LSA) [15] whereby semantic relationships are identified through observing commonality of words across documents in a corpus. Such approaches are limited to verbatim extraction, which limits the fluidity of outputs produced. Modern NLP methods primarily focus on abstractive summarization using sequence-to-sequence deep learning models [22] [18], which result in summaries that often contain new language not present in the original text.

Rapid progress has been made in recent years improving encoder-decoder architectures for text summarization. Convolutional or recurrent neural networks such as the sequence-to-sequence method proposed by Sutskever, Vinyals, and Le [22] have greatly imporoved on the state-of-the-art at the time, while the attention mechanism-based Transformer architectures [23] have quickly risen to prominence for their performance and scalability. The Transformer-based GPT-2 released by OpenAI [17] has proven the performance of unconditional language models on summarization tasks [21], and hybrid approaches such as BART [10] have sought to merge the bi-directional encoder approach of BERT [4] with the left-to-right decoder approach of GPT [16].

While model development for text summarization has rapidly progressed in recent years, training data for summarization has largely been limited to news datasets such DUC, Gigaword, CNN / Daily Mail, and NEWSROOM, with some less popular non-news datasets such as those sourced from knowledge bases [9] and patent records [19]. While summarization is highly relevant in specialized domains such as medicine [12] and finance, it is difficult to find open, specialized and sufficiently large annotated datasets for fine-tuning of pretrained models. While previous literature has discussed the use of Reddit data for text summarization [8] [24], the task of creating corpora specific to particular topics and communities represented by a subreddit or group of subreddits is a novel area addressed through our work.

3. Dataset Preparation

Our dataset is constructed by scraping and cleaning data from the website Reddit which has a number of key features that make it an ideal source of data for text summarization:

- The Reddit upvote system provides a natural filtering mechanism that allows us to remove poor quality posts, generating a high-quality dataset for training
- Users often provide summaries of their posts and of others, indicating the summary with a "TLDR" label
- Wide variety of content with over 1 million subreddits [25], each of which corresponds to a distinct topic

In particular, the vast number of topics gives us the ability to build text summarizers for a wide variety of different domain specializations. We illustrate the benefits of domain specialization by focusing on the finance and medical domains. For finance, we curated a collection of subreddits (table X), and for the medical domain, we focus on the r/AskDocs subreddit, where patients often summarize their question to physicians in a TLDR.

We build upon the work from[25] enhancing it in a number of ways to increase the precision and recall of the summaries being extracted.

3.1. Dataset Collection

We develop one end-to-end script that can be used to scrape Reddit, extract the TLDRs, and write results to a database for further filtering and cleaning. The TLDR extraction and scrapping of Reddit explained below are all present in said script and the python modules that script calls. The final steps in the process of dataset preparation are cleansing and filtering which take place in PostgreSQL scripts.

For the purposes of this paper we focus exclusively on collecting self-posts which correspond to the original posts in Reddit, not the comments. Unlike the work from [25], we exclude comments to maximize the quality of our summaries, as self-posts tend to be longer and include the highest quality TLDRs.

We identify self posts initially by searching for posts that contain both "TL" and "DR". In addition, we filter to the score (upvotes - downvotes) to be above 1 in order to remove lower quality posts. We also include functionality in our scraper to pull from an explicit collection of subreddits specified by the user.

3.2. Dataset Cleaning

We use the code from [25] as our baseline for summary extraction with several key improvements to increase the quality and number of extractions:

- Rather than match TLDRs based on `tl.{0,3}dr` (allowing for up to 3 wild card characters between the "tl" and "dr" strings), we match based on `tl.{0,1}dr`, as we observe that the `tl.{0,3}dr` matching function produces many false positives. Most of these false positives originate from words containing "tl" near the end of the word (e.g. "abruptly") and the next word beginning with "dr".
- Instead of assuming the TLDR match occurs at the end if it is not matched at the very first token, we handle cases when the match occurs elsewhere in the body of the post. This removes the issue of creating massive extracted summaries which happen when a full post is mislabeled as a TLDR.

  This has the added benefit of significantly increasing the size of our training dataset, as TLDRs are often present in the beginning of the post, especially for some of the domain-specialized subreddits we focused on.

- We apply additional filtering of the extracted summaries by removing content examples where
  - summary has fewer than 6 words
  - summary has equal to or fewer words than the content
  - content or summary is not in English
  - author is undefined
  - post was produced by a bot
  - post is deemed to be duplicate after removing non-alphabetical characters and converting all text to lower case.

3.3. Dataset Statistics

Prior to applying our additional filtering described above, our dataset includes 4.1 million summary and content pairs. After applying all of the filtering layers, the cleaned dataset consists of 1,687,257 training pairs. Table 1 provides a summary of key domain-specific subreddits, as well as overall data volumes, in our final dataset.

Table 1: Training Dataset Volumes

Dataset	**Number of Summary Pairs**	**Key Subreddits**
Full	1,687,257	r/relationships (26%), r/tifu (3%), r/AskReddit (2%), r/leagueoflegends (2%), r/trees (1%), r/legaladvice (0.7%)
Finance	3,295	r/wallstreetbets (33%), r/investing (19%)
Medical	3,886	r/AskDocs (100%)

After filtering the posts in the general population, the average length of a summary is 35.52 words and the average length of content is 406.91 words. Some other key statistics for the full dataset and groupings of specialized subreddits can be found in Table 2.

Table 2: Additional Dataset Statistics

Dataset	**Average Words in Content**	**Average Words in Summary**	**Median Words in Summary**
Full	406.91	35.52	25
Finance	365.80	33.53	24
Medical	411.48	39.53	28

4. Model and Experiments

Here we provide a summary of the models used and experiments conducted in order to determine the optimal combination of training datasets and fine-tuning methodologies, provide a summary of our results, and outline and justify the optimal fine-tuning approach.

4.1. Model Architecture

We adopt the BART architecture as the initial pre-trained model, and fine-tune it using the process proposed in the original paper [10]. The BART model is built using the standard Transformer-based sequence-to-sequence architecture [23], and takes a denoising pre-training approach similar to BERT [4], utilizing more aggressive masking and noise functions than the original BERT pre-training. Similarly to the large version of the base model, we use 12 layers for each of the encoder and decoder, respectively, with each layer performing additional cross-attention over the final hidden state of the encoder.

4.2. Experiments

Our main goal is to quantify the benefit of fine-tuning to inject additional domain knowledge into a summarizer that combines extractive and abstractive summarization techniques. To accomplish this we perform 4 primary experiments. In each set of experiments, we evaluate the benefits of pre-training on different subsets of our data and using different fine-tuning approaches, using specialized test sets for the finance and medical by us taken us subsets of training data described in Table 1. Each dataset was randomly split into train, validation, and test sets using 60/20/20 allocations into each.

1. In the first experiment we apply the out-of-the-box BART summarizer that was trained explicitly on the CNN/Daily Mail datasets to our specialized test datasets directly.
2. In the second experiment, we begin with the out-of-the-box BART summarizer, and use the full Reddit dataset for fine-tuning. In this experiment, we only train the decoder, while keeping all the encoder layers frozen.
3. In the third experiment, we follow everything that was done in the second but additionally train the encoder layers.
4. In our final set of experiments, we take the out-of-the-box model and fine-tune it directly on each domain specific training dataset (Table 1), and fine-tune both the encoder and the decoder.

4.3. Fine-Tuning Approach

As is done in BART and initially in RoBERTa [11] we aim for very large batch sizes and accomplish this using gradient accumulation. We hypothesize that large batch sizes are of particular importance for Reddit data due to the noise in this dataset. Managing noise is a challenge common to training summarization models, which often causes peak performance to be achieved very early in the training process. In most of our experiments, we also see imporvement only for the first few epochs of fine-tuning. This is consistent whether fine-tuning on the entire Reddit dataset, or one of our specialized subsets. We expect this is due to the better quality summaries that are present in the CNN/Daily Mail dataset, and the high quality of the out-of-the-box language model. During training we set the maximum number of attempts for validation loss to increase equal to 1 as our stopping criterion. We do this to conserve computational resources, as due to our very large batch size the validation loss steadily decreases after it initially starts going down.

For autoregressive generation parameters we match the original BART paper where appropriate, matching their length penalty, number of beams, minimum number of output tokens, and removal of duplicated tri-grams. We set the maximum number of output tokens to 120 (resulting in average output length of approx. 80 words for all of our models), maximum number of input tokens at 512, and use nucleus sampling [7] with Nucleus p = 0.95 as is proposed in the paper. For additional details of hyperparameters and their tuning refer to Table 3.

A cluster with one NVIDIA Titan XP GPU, 8 vCPUs and 12 GB of host memory is used to train the models. We expect that by scaling GPU memory and increasing model size, e.g. through raising the maximum sequence length to account for longer-term dependencies in the input.

Table 3: Hyperparameter Selection

Hyperparameter	Final Value Chosen	Indicative Range Tested
Learning Rate	$1e^{-4}$	$1e^{-4}/1e^{-5}/1e^{-6}$
L2-Lambda	0.01	0.001/0.01/0.1/1.0
Length Penalty	1.0	1.0/2.0/10.0
Batch Size	1024	512/1024/8000
Number of Beams	5	N/A
Repetition Penalty	2.5	N/A
Max Output Tokens	120	N/A
Min Output Tokens	56	N/A
Max Input Tokens	512	128/256/512
Nucleus P	0.95	N/A
Learning Rate Scaling	0.95	N/A

5. Results

We find that the highest ROUGE scores on our test dataset are obtained using the fine-tuning approach in our fourth set of experiments, where we fine-tune on the specialized datasets directly.

We find that for training the model on the full corpus that freezing the encoder led to a significant improvement in the model performance on each domain, resulting in a much better performance on our test set in the second set of experiments compared to the third set of experiments (see Table 4). For the domain-specific summarizers, no layer freezing led to the best outcomes. We believe this indicates that allowing for encoder adaptation to the vocabulary can lead to additional improvements when training on a new specialized domain.

5.1. Model Evaluation

We use ROUGE-1 recall, ROUGE-2 recall, ROUGE-1 precision and ROUGE-2 precision as our primary evaluation metrics. In order to generate the summaries, we use teacher forcing for training and autoregressive generation for evaluation. We generally match the BART architecture with most of our experiments focused on optimizing fine-tuning. We experimented with freezing different sections when training both for the encoder and the decoder to control learning rates across layers, and find that different training tasks performed best with different layer freezing setup.

Table 4 summarizes our experimental results. We hypothesize that the improved performance from encoder freezing for the full corpus model and not the domain specific models is due to the quality and type of the summaries that are most common in the full corpus. In the full corpus most of our summaries are from subreddits that use very different language than our domain-specific ones, potentially making the encoder fine-tuning counterproductive; . The models fine-tuned on all of Reddit data without any layer freezing (experiment 3) did not perform significantly better than the out-of-the-box pre-trained model on specialized domains. Significant gains on the domain specific datasets are achieved from training on the general dataset with the encoder frozen, and even further gains from fine-tuning on each specific domain.

We find reddit data alone to be insufficient to train a full summarization model. This is clear when considering how heterogeneous language, as well as format and quality of summaries, are across subreddits, making it a less attractive dataset for training a base summarizer than traditional datasets like CNN / Daily Mail. This makes fine-tuning a pre-trained summarizer using our domain specialized datasets the better strategy compared to training a specialized summarizer from scratch.

As the corresponding ROUGE scores in Table 4 indicate, summarization quality improves greatly as more specialized data is used for training summarizers. This can also be seen qualitatively, with the more

Table 4: Experiment Results

Model	**R1**	**R2**	**P1**	**P2**
1. Train - CNN/DM	31.37	3.62	11.18	1.19
2. Train - CNN/DM; Fine-tune - General; Encoder Frozen	35.70	**4.79**	**12.39**	**1.57**
3. Train - CNN/DM; Fine-tune - General; Fully Unfrozen	31.69	3.76	11.93	1.34
4. Train - CNN/DM; Fine-tune - Finance; Fully Unfrozen	**36.94**	4.54	12.18	1.42

specialized models learning to tailor their summaries more to the specific domain. As we discuss in the qualitative observations section below, the structure over the Reddit TLDR summaries also offers some advantages over the CNN / Daily Mail dataset, which in many cases leads to potentially more useful summaries.

5.2. Qualitative Observations

We give some examples of summaries generated by the best models from experiments 1,3, and 4 (omitting the results of experiment 2 where models fine-tuned on the full reddit data with both the encoder and decoder trained). We also omit the original text of summaries for brevity and to protect Reddit users' privacy, but discuss the quality of each summary separately.

In both of the examples in Table 5 the specialized fine-tuning model picks out relevant medical facts from the original summary. Interestingly, the specialized summarizer also picks up on the fact that the chest pain was specifically not a heart attack, which both of the other summarizers fail to do. We see a similar pattern from the second example in Table 5, where the medical specific summarizer extracts the fact that in the follow up session the doctor can do additional imaging or a biopsy on the lesion, which the first two summarizers also fail to do.

Table 6 we see some examples from where the specialized summarizer creates summaries similar in language to that used in r/wallstreetbets, one of the most prevalent finance subreddits in our corpus. This serves as an illustration for some caveats of using Reddit data for NLP domain adaptation.

Although both the General Reddit and Specialized model are able to capture a lot of the financial context of the underlying post, the finance example also showcases some of the pitfalls of domain specialization. Given the prevalence of r/wallstreetbets data in our test set, the specialized model achieves higher ROUGE scores by biasing towards much of the colloquial langauge present on that subreddit, and less so on finance specific text. Nevertheless, both of these models perform much better qualitatively on the financial context of text data than the out-of-the-box CNN / Daily Mail model.

Overall, we see a number of qualitative advantages for our fine-tuned models across both of our specialized domains. One common theme from the out-of-the-box model is that it often ends in recommending a URL, sending an email to someone, or sharing a perspectives on social media. Qualitatively, the out-of-the-box summarizer also focuses a lot more on the narrative, generating summaries that are well-written and appear intriguing, but omit key facts, which can also be seen in Table 5 and Table 6. We believe the Reddit summarizers extract these additional details more effectively as the training summaries from Reddit are better designed to do so, as they are intended to serve as a replacement for reading the post, rather than a ”hook” to get the reader to engage with the content.

Table 5: Sample Results - Health Summarizer

Out-of-the-Box Model	**General Reddit Fine-Tuning**	**Specialized Fine-Tuning**
Had some chest pain last night but things don't add up. No longer experiencing this issue, curious what it might have been ¬†and how to find out more about Ehlers-Danlos and other conditions that may cause the same symptoms in a person of similar age or different health issues than those I am currently suffering from! Have you had any unusual experiences with your body? If so please share them on our Facebook page at: `http://soulmatestories@dailymailonline`	I'm a 24 year old Caucasian female BMI of 20115 pounds and 5'3.5 I have asthma, Ehlers-Danlos or some other condition that causes chest pain? If it wasn't an issue in the past can you tell me what might be going on with my heart attack last night - It's not like there is any clotting risk factors aside from Nuvaring for birth control so why did this happen after dancing around all day long to no one else who	had pain in my chest and I am sure this couldn't have been a heart attack, what if it was something else that would make me feel better? What do you think is going on with the tachycardia around 115 BPM or should be an issue for someone who has no clotting risk factors aside from Nuvaring to worry about after dancing last night there's not enough blood pressure readings so maybe its just low sodium levels because of being at work every 30 minutes - can
Liver MRI-Focal Nodular Hyperplasia or something else? Hello docs of reddit! I'm hoping you guys can help me out. A couple months ago i went to the doctor for my psoriatic arthritis diagnosis and a ct scan showed two liver lesions that are nonspecific, but definitely not hemangiomas¬†probably F focal nd narcolepsy/hyperparathyroidism (HPC) The doctors suggested follow up imaging with either	Liver MRI-Focal Nodular Hyperplasia or something else? I'm not sure if some prior info would be helpful to you guys. A couple months ago got my psoriatic arthritis diagnosis and am still in a lot of pain, but don't want the biopsy just because it's possible that its Hepary Adenoma/Hemangiomas - can someone help me out with this one please! If anyone has any information about liver imaging for people who	I I'm not sure if some prior info would be helpful to you guys? liver MRI-Focal Nodular Hyperplasia or something else that my doctor can do follow up imaging and a biopsy on the 2.1 cm lesion along with an unmeasured but smaller lesions are nonspecific, maybe it's Hepatic Adenoma for another child soon - don't want all these tests just because of what they think is going in there probably no reason why this should

6. Significance and Future Work

Our research demonstrates the potential application of our data extraction and fine-tuning pipeline in the fields of finance and healthcare. Models fine-tuned in a manner described in our paper can be used as standalone text summarization models, or further improved by augmenting training data with additional public or proprietary data sources. More importantly, we show a clear positive trend in the quality of text summarization for a large language model such as BART when fine-tuned using specialized data, even in cases where the amount of labeled examples available for fine-tuning is relatively small. We also quantify the relative extent to which additional specialization can improve the quality of summarization. We hope these data points can be used by industry practitioners to help with the cost-benefit analysis for selecting the right level of fine-tuning, from using large out-of-the-box training models to obtain potentially costly high-quality specialized data.

Based on the qualitative evidence of domain specialization and the increase in the ROUGE-1 scores from specialization over the general model with the encoder frozen we conclude that the domain specific models are able to specialize into their respective domains. The models do this by learning what parts of the text are more significant based on the domain, and learning to more effectively leverage relevant

Table 6: Sample Results - Finance Summarizer

Out-of-the-Box Model	**General Reddit Fine-Tuning**	**Specialized Fine-Tuning**
The efficient market hypothesis supposes that an individual can't beat the markets. Most advice floating around rfinance andrinvesting is with regards to long-term investing through a range of ETFs, writes John Defteriosa in his column for The Motley Fool's "Moneyquiz" series on Sunday nights at 10pm ET/PT (11:00 pm BST) If most people believe in this strategy then why haven't more shrewd investors figured out how	Question ETFs and efficient markets. Why haven't the more shrewd investors figured out a way to essentially capture this future growth via arbitrage? If most people believe in that strategy, why isn't all of it priced-in into these investment strategies already at risk-reward trade off? The idea is an investor enjoys full diversification and low fees on their investments through exchange traded funds (ETF's) What do you think about whether or not there are enough opportunities	why aren't all of the growth priced in to these ETFs already? Why is this a good strategy, and isn't arbitrage better than an individual investing through them on long-term risk-reward trade off. If I can buy my way into it now would be great for me but not you or your investor if they think that's what everyone else does as well - by going with short term returns over 50% market return + low fees will go both ways!

vocabulary when developing abstractive summaries. Due to these observations we believe domain adaptation capabilities are an excellent use of the dataset we have created. Those who use this dataset for the purpose of domain-specialization should make sure the language and discourse in these subreddits aligns well with the business problem at hand.

In addition to the domain-specialization use case, due to the increase in the performance and summary quality on each domain from training on the entire Reddit dataset with encoder frozen, we also believe there is value in using this dataset to improve general summarization models. Those interested in using this dataset for general summarization should note the importance of layer freezing to avoid the model picking up unwanted language and patterns from general Reddit data.

The summaries dataset can be extended in a number of ways to provide additional value. There is a large number of TLDRs that summarize website content linked to in Reddit posts, and scraping the content of these website can augment the summaries dataset with high-quality data. Summaries can also be augmented by parsing through comment trees and extracting TLDRs from long comments.

The summarization models can also be improved in several ways. A high-potential area of improvement is controlling catastrophic forgetting, via regularization schemes such as layer freezing or varying learning rates across layers. Forgetting can also be addressed through modifying training data, by augmenting fine-tuning data with samples of the original summaries. Finally, new architectures like XLNet [28] can be employed to increase the max sequence length the transformer encoder-decoder can handle to improve performance on very long documents.

References

[1] Mehdi Allahyari et al. *Text Summarization Techniques: A Brief Survey.* 2017. arXiv: 1707.02268 [cs.CL].

[2] Sumit Chopra, Michael Auli, and Alexander M. Rush. "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* San Diego, California: Association for Computational Linguistics, June 2016, pp. 93–98. DOI: 10.18653/v1/N16-1012. URL: https://www.aclweb.org/anthology/N16-1012.

[3] Arman Cohan et al. *A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents.* 2018. arXiv: 1804.05685 [cs.CL].

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* 2018. arXiv: 1810.04805 [cs.CL].

[5] Max Grusky, Mor Naaman, and Yoav Artzi. *Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies.* 2018. arXiv: 1804.11283 [cs.CL].

[6] Donna Harman and Paul Over. "The Effects of Human Variation in DUC Summarization Evaluation". In: *Text Summarization Branches Out.* Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 10–17. URL: https://www.aclweb.org/anthology/W04-1003.

[7] Ari Holtzman et al. "The Curious Case of Neural Text Degeneration". In: *ArXiv* abs/1904.09751 (2020).

[8] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. *Abstractive Summarization of Reddit Posts with Multi-level Memory Networks.* 2018. arXiv: 1811.00783 [cs.CL].

[9] Mahnaz Koupaee and William Yang Wang. *WikiHow: A Large Scale Text Summarization Dataset.* 2018. arXiv: 1810.09305 [cs.CL].

[10] Mike Lewis et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.* 2019. arXiv: 1910.13461 [cs.CL].

[11] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[12] Rashmi Mishra et al. "Text summarization in the biomedical domain: a systematic review of recent research". In: *Journal of biomedical informatics* 52 (2014), pp. 457–467.

[13] Ramesh Nallapati et al. *Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond.* 2016. arXiv: 1602.06023 [cs.CL].

[14] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. "Annotated Gigaword". In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction.* AKBC-WEKEX '12. Montreal, Canada: Association for Computational Linguistics, 2012, pp. 95–100.

[15] Makbule Ozsoy, Ferda Alpaslan, and Ilyas Cicekli. "Text summarization using Latent Semantic Analysis". In: *J. Information Science* 37 (Aug. 2011), pp. 405–417. DOI: 10.1177/0165551511408848.

[16] Alec Radford et al. "Improving language understanding by generative pre-training". In: *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf* (2018).

[17] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: 2019.

[18] Alexander M. Rush, Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization". In: *CoRR* abs/1509.00685 (2015). arXiv: 1509.00685. URL: http://arxiv.org/abs/1509.00685.

[19] Eva Sharma, Chen Li, and Lu Wang. *BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization.* 2019. arXiv: 1906.03741 [cs.CL].

[20] Tian Shi et al. *Neural Abstractive Text Summarization with Sequence-to-Sequence Models: A Survey.* 2018. arXiv: 1812.02303 [cs.CL].

[21] Sandeep Subramanian et al. *On Extractive and Abstractive Neural Document Summarization with Transformer Language Models.* 2019. arXiv: 1909.03186 [cs.CL].

[22] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. *Sequence to Sequence Learning with Neural Networks.* 2014. arXiv: 1409.3215 [cs.CL].

[23] Ashish Vaswani et al. *Attention Is All You Need.* 2017. arXiv: 1706.03762 [cs.CL].

[24] Michael Völske et al. "TL;DR: Mining Reddit to Learn Automatic Summarization". In: Jan. 2017, pp. 59–63. DOI: 10.18653/v1/W17-4508.

[25] Michael Völske et al. “TL;DR: Mining Reddit to Learn Automatic Summarization”. In: Jan. 2017, pp. 59–63. DOI: 10.18653/v1/W17-4508.

[26] Lucy Lu Wang et al. “CORD-19: The Covid-19 Open Research Dataset”. In: *ArXiv* (2020).

[27] Robin Wigglesworth. “How investment analysts became data miners”. In: *Financial Times* (Nov. 28, 2019). URL: https://www.ft.com/content/8d3aaf42-108b-11ea-a7e6-62bf4f9e548a.

[28] Zhilin Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *NeurIPS*. 2019.