The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Interactive Summarization for Data Filtering and Triage

Justus Robertson
University of York
York, United Kingdom
justus.robertson@york.ac.uk

Brent Harrison
University of Kentucky
Lexington, KY 40508
harrison@cs.uky.edu

Arnav Jhala
North Carolina State University
Raleigh, NC 27606
ahjhala@ncsu.edu

Abstract

There is an increasing demand for content filtering and flagging on social media in relation to cybersecurity and social media conduct monitoring. This task is challenging and there is a large body of recent work that addresses it within the Natural Language and Video Processing communities. In this work, we propose two novel perspectives on this problem and provide preliminary evidence for their potential success. First, for text-based data, we utilize the current state of the art topic-based summarization algorithms and provide an interactive topic-conditioning approach to enable multiple summarizations based on different highlighted topics. Second, due to the interactivity aspect, we are able to characterize how this approach can be integrated within the process of a human analyst to improve both the quality of filtered data and the effort.

Introduction

Every day, we produce large amounts of data relating to many aspects of our everyday lives. By some estimates, we generate 2.5 quintillion bytes of data daily spread over social media platforms such as YouTube, Facebook, and Twitter. This wealth of data is often of great interest to analysts and researchers in a variety of fields such as cybersecurity and marketing. While the size and scope of this data makes it interesting for analysts and researchers, this also makes it difficult to work with. In order to derive any actionable insights from such a large quantity of data, analysts must be able to effectively filter and triage data in order to ensure that only relevant data is retrieved.

Data filtering and triage refers to a problem that is common in the analysis community in which analysts and researchers must downselect information from a large corpus of data such that only relevant information is retrieved. This may involve returning documents relating to a specific topic from amongst a large dataset of documents. It may also involve selecting only elements of a single document that are relevant to a search query.

Given the size of data that analysts and researchers often work with, this is a very difficult and often time consuming task to perform. Often, those performing these tasks develop workflows that are based on intuition, past experience, and technical expertise. Thus, it can be difficult for those new to data filtering and triage to effectively sort through large amounts of data to perform this task as they have not had the time to build up this expertise and intuition. In these situations, an artificial intelligence or machine learning system could provide assistance by performing in initial filtering such that the analyst or researcher only has a smaller amount of data to filter and triage.

Current systems that are used by analysts utilize hand authored rules for navigating and filtering datasets. Query construction based on analyst goals for investigation into patterns of data is cumbersome and time consuming both in design and usage. Rigorous characterization of the process or the integration of NLP tools in the process in terms of effectiveness of human-machine collaboration is lacking in current research. Better understanding of the rules that are explicitly or implicitly used by human analysts in forming queries need to be learned. By learning these rules (either explicitly or implicitly) through machine learning, we expect to lower the authorial burden and cognitive load on the analyst, which makes this a more approachable method for performing filtering and triage. One set of techniques that have not been readily explored for this task are techniques related to automatic summarization. We feel that summarization can be viewed as a means to quickly convey important information to a reader, which is one of the core necessities when designing a data filtering and triage system. In this paper, we will discuss how current summarization techniques could be applied to data filtering and triage problems and how current approaches address unique issues in this area. We also identify a set of future research areas to be explored to improve the applicability of current summarization techniques to data filtering and triage.

The remainder of the paper is organized as follows. In section 2, we will more formally define the data filtering and triage problem. In Section 3 we will discuss how current automatic summarization techniques could be applied to the data filtering and triage task using the desiderata outlined earlier within an interactive system to provide support to users. In section 4 we will discuss some of the insights

and lessons learned via our case study on topic-based summarization with the CNN / Daily mail dataset. In Section 5 we will discuss potential ways that automatic summarization techniques can be used to further support data filtering and triage tasks. We also identify some potential future avenues of research in this area. Section 6 provides an overview of related work both in summarization and topic modeling that is relevant to this task.

Data Filtering and Triage

To provide a specific use case for the type of summarization we are addressing in this work, this section provides an overview of the process for a typical analyst. More specifically, we describe the process of a cybersecurity analyst or social media moderator who is observing a stream of public social media data and attempting to keep track of threads of potentially inappropriate content. The process of filtering content to support an analysis could potentially introduce significant errors and biases, and so methods and tools for data triage must be characterized in a way that makes clear their appropriate use in rigorous analytic processes. There are two types of tasks that are performed. First, formulating questions after an event and looking for patterns of past posts that could potentially be useful in predicting future similar events. Second, keeping track of streaming data by tagging it. These tasks can be broken down in terms of requirements for computational tools in the following way. We will note here that this paper does not attempt to propose solutions to the entire process but a full description is presented along with challenges for a complete overview of the end-to-end process.

- **Data Retention :** One key challenge in continuously monitoring incoming data in various forms is to identify and retain data that is most likely to be useful in the analysis process. One way to address this problem is to store more semantically compacted data in the form of appropriate summaries. Another way is to find relationships between multiple sources of information on the same topic and retain key information.
- **Search/Filtering :** The next challenge is in providing interfaces to underlying computational methods and algorithms that could support various tasks such as search, filtering, aggregating, determining relationships between information coming from disparate sources, and formulating refined queries for identification of relevant data sources to drive the search/filtering process.
- **Prioritization :** Search and filtering process results in a ranked list of relevant data elements pertaining to the query and query parameters. From the results of a number of queries and query parameters, analysts need to quickly prioritize on key topics to escalate or reformulate more precise queries.
- **Presentation/Collaboration :** Finally, analysts need to summarize, discuss, and collaborate on escalated posts or topics for further monitoring and present these results through clear visualizations and justifications.

Each of these tasks has their own set of unique challenges associated with them, and providing a detailed discussion about how automatic summarization techniques can be used to address them is beyond the scope of this paper. Instead, we focus on the data retention task. Specifically, we explore how existing summarization techniques can be adapted to this task by having them perform topic driven summarization and the lessons learned throughout this process.

Initial Steps: Topic-Driven Summarization for Data Retention

Earlier we mentioned several challenges that must be addressed when performing data filtering and triage. In this paper, we present a case study of our initial steps in using automatic summarization techniques for this task. Specifically, we outline how we extend current methods to make them more applicable for performing the data retention task by enabling them to tailor news article summaries based on a given topic.

Narrative text, like news articles, often have many different interrelated topics. For example, an article about the 2019 US Women's National Soccer Team may include information about international sporting events, US politics, and the FIFA organization. There are many possible ways to summarize text to convey different perspectives and nuances in the original source. Accounting for topic is one way to direct automated summarization systems to tailor their output for human users. This section describes initial steps towards a topic-driven abstractive summarization system.

Dataset Construction

A high level overview of our dataset creation pipeline is given in Figure 1. We start with the non-anonymized CNN / Daily Mail dataset (See, Liu, and Manning 2017) which contains a total of 300,000 online news article and multi-sentence summary pairs. To learn the relationship between articles, summaries, and topics, we must also have topic information for each summary. Since this information is not included with the CNN / Daily Mail dataset, we must find some way to assign topic information to this dataset ourselves. To do this, we use Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004) to train a topic model of the summaries in the original corpus. Our model includes 147 topics which are clusters of related words assigned to an integer. For example, here are the first three topic clusters in our model:

> Topic 0: apple phone new mobile iphone phones app available devices ipad launch screen device google samsung android expected users company use microsoft version tablet watch service
>
> Topic 1: cent people survey percent study half average likely shows poll 000 uk according say americans 10 just finds 50 research report nearly 40 adults 80
>
> Topic 2: weight size lost stone diet pounds lose body dropped fat weighed loss fitness exercise eating just 12 healthy food day fit 10 weighs months gym

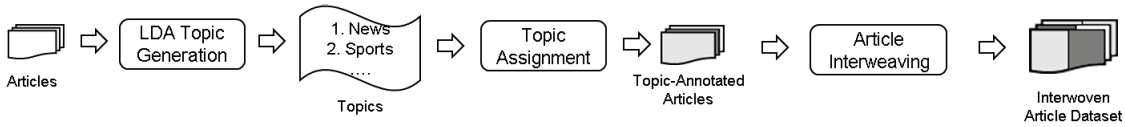

Figure 1: Overview of our training set creation pipeline.

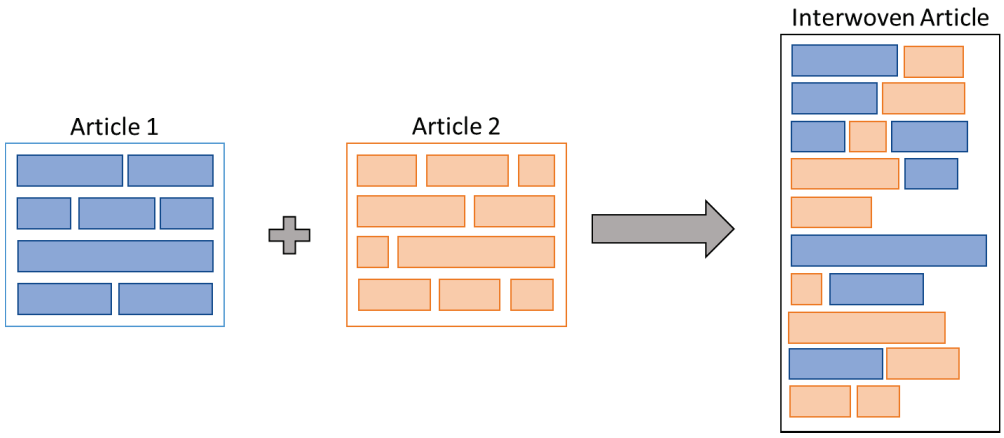

Figure 2: A high level overview of how we created interwoven articles for training. Boxes in this figure represent sentences. Ordered pairs of sentences were taken from two parent articles and inserted into the interwoven article.

Topic 0 corresponds to mobile and tablet tech devices and companies, Topic 1 to scientific studies and public opinion polls, and Topic 2 to fitness and weight loss. Here, topics consist of words that are highly related to each topic. Thus, each topic is, in essence, a bag of words that we can use to assign topics to each article in our dataset.

Once the topic model has been trained, we apply the topics to each summary in the original corpus and find the probability that the summary belongs to each of the topics in our model. For example, the summary:

> Once a super typhoon, Maysak is now a tropical storm with 70 mph winds. It could still cause flooding, landslides and other problems in the Philippines.

Would be assigned the topic:

> Topic 17: new people storm hit homes power residents damage hurricane officials 000 says area flooding say quake miles reported earthquake struck dead toll tornado caused flood

The model assigns this topic a probability of 85.6%. This probability is generated by looking at the words in the summary and comparing them to the words in the topic cluster. Here, notice that several words, such as "flooding" and "storm," show up in both the topic cluster and the summary. This indicates that it is likely that this summary concerns this topic.

Once all summaries have been assigned, we follow the procedure in (Krishna and Srinivasan 2018) to randomly merge pairs of articles and associate them with each of the original summaries. Specifically, this involves constructing a new article by taking a sentence from each article and interweaving them to create a new article. So, the first and second sentences in this new interwoven article would be the first sentence from each of the parent articles. This process continues until each sentence has been used from both articles. An overview of this process can be seen in Figure 2. Since the clustering that we performed earlier assigns a topic to each article, this procedure ensures that each interwoven article contains multiple topics. This also ensures a strong coupling between each sentence and a topic. This will help force the summarization model to learn which portions of text are important to summarize for each topic.

Models, Training, and Results

Once we have created this initial training dataset, we use it to train an automatic summarization system. We start with a baseline pointer generator network (See, Liu, and Manning 2017) which is an augmented neural sequence-to-sequence abstractive summarization architecture. One of the primary features of a pointer generator network is their ability to reuse input text instead of having to generate novel text. This makes them an ideal architecture for performing summarization. We train the model for 180,000 iterations on our modified CNN/Daily Mail dataset augmented with summary topics. The goal of this baseline is to see how a baseline pointer generator would attempt to summarize an article in which there were multiple topics present.

We hypothesize that a baseline network will struggle with this task because there are, in essence, two signal sources in each document. Without a way to distinguish between each signal, the network will be forced to choose to generate a

summary based on one of these signals. In other words, in a document that contains articles concerning sports and politics, we expect the baseline network will choose to summarize either the sports part of the article, or the politics part of the article, but not both. To help the network better identify what information to summarize, we also train a second pointer generator network where the summary topic integer is prepended to the usual news article in the model's input. We expect the topic information model to produce more accurate summaries as measured by ROUGE metrics since this will allow it to learn relationships between topics and summaries. The results after training the second model for 90,000 iterations are shown in Table 1.

Our results confirm our hypothesis. The network with topic information outperforms the baseline network on each of the ROUGE metrics. Here is an example summary from the corpus about a trial run of robotic employees:

> Robot will work on a trial basis at Mitsubishi UFJ Financial Group branches. These trials of the 1ft 11inch 58cm assistant are expected to begin in April. Nao has four microphones, touch sensors and can speak 19 languages. Makers Aldebaran Robotics said it can also recognise human emotions. If successful, the robotic employees will be rolled out to more branches.

Each article is intermixed with a second article, assigned each of the original summaries, along with their LDA topic cluster. Here is the summary of the second article, randomly selected to be intermixed with the first about robots. It is about a young man suspended from school for selling sodas to fellow classmates:

> Grade 12 student Keenan Shaw, 17, was handed a two-day suspension. He was told the sales violated the school 's nutrition and marketing policies and that he was operating a business without a licence. Keenan defended actions by pointing out other students have been known to sell marijuana, cigarettes, acid and even meth. Has now moved his business outside school to sidewalk.

Each summary is associated with a topic by LDA in the dataset. In this case, the first summary's topic is related to technology and design:

> Topic 143: used using technology device uses created designed use design 3d developed make machine light create company called project robot built able computer machines firm sensors

Trained on just the intermixed article and summary pairs, the baseline model produces the following summary:

> 12 Keenan,,,, from, week after was to selling at profit his locker. Was a suspension Winston High in,,,, selling. School 's nutritional policy, sodas not in.

Without topic information, the model summarizes the wrong intermixed article and produces strange grammar and punctuation errors. Here is the summary produced by the network trained with topic information:

> Experts have warned robots could soon take over our jobs. It has two cameras mounted to its head, that act as eyes, as well as four directional microphones to act as its ears.

Given topic 143, the second model correctly identifies and summarizes information from the intermixed article that corresponds to technology. This difference in output accounts for the stronger ROUGE scores by the topic-driven model.

Discussion and Lessons Learned

We have shown how a slight alteration to the input of a pointer generator network can better enable them to generate topic-centric summarizations in the case where an input document contains text concerning multiple topics. Through the course of this study, however, we learned many lessons concerning model and dataset creation that could be of use to researchers who want to pursue this line of research. We will discuss these in greater detail below.

The Need for Datasets

The first difficulty that we encountered was the lack of a suitable dataset for this task. While there are many large naturally occurring text corpora that exist, we were unable to find any that contained granular topic information for single documents. Often, topic information is either given on a document level or, as is the case with the CNN daily mail corpus, not at all. While these datasets that do specify topics often contain documents that are associated with multiple topics, they do not tightly couple sentences or groups of sentences in the document with this topic information. It was due to this fact that we had to synthetically construct an *interleaved* topic dataset so that we could evaluate how well our summarization techniques were able to extract information relevant to a query topic.

While we feel this was a suitable direction to take for this work, we do not see this as the best way to proceed moving forward. By interleaving news articles we ensure that the resulting article is composed of multiple topics; however, the transitions between topics are abrupt, which potentially makes it easier for a machine learning model to identify them. Moving forward with this research will require specialized datasets that identify topics within documents at a sentence, or group of sentences, level.

The Need for Robust Topic Representations

Our initial exploration into topic-based summarization yielded positive results by augmenting the input to a pointer generator network to include topic information. We showed that this augmentation, small as it may be, was enough for the network to begin to distinguish between the text associated with each topic in our test examples. While these initial results are positive, we feel that this may largely be because of the nature of our dataset and, as such, more robust topic representations will need to be explored in the future. Recall that our dataset consisted of two articles that were interwoven to create one article to ensure that each article contained multiple identifiable topics. Since we identified topics using LDA, each topic essentially consisted of a different bag

Model	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	16.23	1.23	14.84
Topic-Driven	26.02	9.15	23.43

Table 1: ROUGE results for the baseline and topic-driven model on the mixed CNN / DailyMail corpus.

of words. This makes distinguishing topics an overall easier task. By using more robust topic representations, such as a continuous topic vector (Le and Mikolov 2014), it is possible that the additional information contained within could allow summarization techniques to better distinguish between subtle shifts in topics.

In addition, robust topic representations should enable machine learning systems to better learn relationships between topics. The benefit of using something like a continuous topic vector is that the vectors themselves contain semantic information that can be used to identify relationships between topics. This could aid machine learning models in detecting subtle changes in topics over time or identifying which topics are likely to co-occur in the same document.

Future Directions

We have shown promising results for using automatic summarization techniques for data retention, there are still other aspects of data filtering and triage that we have not explored. In the following sections, we will discuss how automatic summarization techniques could address the other tasks involved with data filtering and triage and the potential for future avenues of research in each one.

Search/Filtering

The primary challenge involved in the search/filtering task is in developing interfaces that provide access to algorithms that support tasks such as data retention. Since this task involves human interaction, the machine learning algorithms that power the data filtering process need to be able to handle a non trivial amount of variability. In the presented method on topic-driven summarization, we augmented the article's input with an input that identified the type of summarization that should be returned. Our dataset contained 147 potential topics. If an interface were to be designed around this system, a mechanism would need to be in place that could either enable users to select which of the 147 topics they wanted to filter by or a way of grounding a free-form user query onto one of these 147 topics. While this is one method of making summarization techniques work well with search and filtering interfaces, it could be problematic if users choose to search based on a topic unknown to the machine learning model. In these cases, having a more robust topic representation can be beneficial. Topic vectors, for example, would allow for users to input any input query to search. This query could then be converted into a topic vector and either directly input into the data retention model or compared against the full set of topics to ground the query onto a pre-existing topic.

Another key feature in search and filtering task is that these interfaces need to be able to retrieve data across a large number of documents. This may involve developing heuristic search techniques that enable the system to quickly identify which documents or articles are likely to contain relevant information and then prioritize them when generating topic-based summaries. This would enable the system to quickly retrieve relevant data across a large number of documents.

Prioritization

Often, analysts and researchers must filter data from large databases. This means that a single query could return thousands, sometimes millions of results. It is likely that many of these results may only be marginally relevant to the researcher's original search query. Thus, a system that aids in data filtering and triage should be able to reason about its own results and identify the results that are most relevant to the current search query. Currently, summarization techniques cannot reason about how relevant a summarization is to a given topic. In this situation, using a bag of words model for defining topics provides a method for determining how relevant a summary is to a given topic. It is possible to use metrics such as word frequency to determine the probability that a summary is associated with a certain topic by comparing against the bag of words for each topic. Thus, the more relevant summaries would have a higher probability of being associated with the query topic.

There is also an opportunity to integrate work on query prediction into topic-based summarization. If topic-based summarization systems could be augmented with temporal reasoning over topic queries, it is possible that they could retrieve and summarize about potential future queries. Many of the current query prediction and generation techniques rely on hierarchical sequence-to-sequence networks that can reason over past inputs and outputs. It is reasonable to assume that a similar strategy could be applied to summarization techniques.

Presentation/Collaboration

Another important aspect of the data filtering and triage process is presenting the data in a concise, yet informative way to analysts and researchers. There are several reasons why this may be important for an automated system. The first of these is that visualizations can help the analyst understand why data may be relevant to the current query or how it fits in with other returned data. Another way that visualizations can be beneficial for this task, especially when working with an automated system, is that visualizations can help the analyst understand why the system is returning certain pieces of data. The ability to understand the decision making process behind automated system is critical for facilitating seamless human-machine cooperation.

There are several interesting ways that data generated by an automatic summarization system can be visualized de-

pending on the overall goal of the visualization. Recall that pointer generator networks have the option to use portions of the input when creating a summary. This gives us the ability to easily ground certain aspects of a summary in its original context. By showing the analyst or researcher this context, they will get additional context about the data returned. This also enables them to discover connections between data that the automated system may not have identified.

Related Work

The CNN / Daily Mail corpus was originally introduced for reading comprehension tasks (Hermann et al. 2015) and was later processed for summarization (Nallapati et al. 2016). These original datasets were anonymized through pre-processing to replace named entities with numbered references. The dataset was later non-anonymized (See, Liu, and Manning 2017) to remove pre-processing requirements. We start from this non-anonymized version of the dataset when constructing our topic-annotated corpus. We process the dataset to add topic information (Krishna and Srinivasan 2018) to the corpus summaries. We extract these topics from the original dataset with LDA (Griffiths and Steyvers 2004).

Since its introduction, many models (Celikyilmaz et al. 2018; Dong et al. 2018; Li et al. 2018; Zhou et al. 2018) have been benchmarked on the non-anonymized CNN / Daily Mail corpus, each making improvements on the scores presented in the original pointer generator model (See, Liu, and Manning 2017). Recently, the pre-trained transformer BERT (Devlin et al. 2019) has been fine-tuned for extractive summarization with strong CNN / Daily Mail results (Liu 2019). For our initial results, we used the earlier pointer generator network (See, Liu, and Manning 2017) and we plan to progress to more recent models as our work continues.

Conclusion

This paper motivates and introduces the problem of integrating NLP tools in the process of data exploration and triage. We have organized the process for this activity to provide a research framework that could be useful for future research and in determination of benchmark datasets. More specifically, we present a use case of security analysts and social media moderators in their process of data capture, filtering, querying, and reporting. We show preliminary work in using state of the art summarization algorithms and present an illustrative example of using topic models to interactively condition the output in a way that is useful to a user interacting with the dataset. Based on this implementation and problem framework, we then present future directions in interactive summarization with a user centered perspective.

As we have shown in this paper, we feel that automatic summarization techniques show great promise for advancing the state of the art on data filtering and triage. While we show promising results in the area, we also acknowledge that there is much more work that needs to be done in this area. We hope that our initial work in this area and our identification of future research areas inspires researchers to consider the broader applications of summarization research with respect to data filtering and triage. We also hope that researchers will consider broader applications of data filtering and triage in the future. While we focus on the needs of analysts and researchers in this paper, data filtering and triage have broader applications in search and data retrieval, two tasks that we perform daily. We feel that these reasons combine to make data filtering and triage an exciting application area for summarization work.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Celikyilmaz, A.; Bosselut, A.; He, X.; and Choi, Y. 2018. Deep Communicating Agents for Abstractive Summarization. In *Annual Conference of the North American Chapter of the ACL*, 1662—-1675.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Annual Conference of the North American Chapter of the ACL*, 4171—-4186.

Dong, Y.; Shen, Y.; Crawford, E.; van Hoof, H.; and Cheung, J. C. K. 2018. BanditSum: Extractive Summarization as a Contextual Bandit. In *Conference on Empirical Methods in Natural Language Processing*, 3739—-3748.

Griffiths, T. L., and Steyvers, M. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228–5235.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems*, 1693–1701.

Krishna, K., and Srinivasan, B. V. 2018. Generating Topic-Oriented Summaries Using Neural Attention. In *Conference of the North American Chapter of the ACL: Human Language Technologies*, 1697–1705.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196.

Li, W.; Xiao, X.; Lyu, Y.; and Wang, Y. 2018. Improving Neural Abstractive Document Summarization with Structural Regularization. In *Conference on Empirical Methods in Natural Language Processing*, 4078–4087.

Liu, Y. 2019. Fine-tune BERT for Extractive Summarization. *arXiv preprint arXiv:1903.10318*.

Nallapati, R.; Zhou, B.; Gulcehre, C.; Xiang, B.; et al. 2016. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In *SIGNLL Conference on Computational Natural Language Learning*.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the Point: Summarization with Pointer-Generator Networks. In *Annual Meeting of the ACL*, 1073—-1083.

Zhou, Q.; Yang, N.; Wei, F.; Huang, S.; Zhou, M.; and Zhao, T. 2018. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In *Annual Meeting of the Association for Computational Linguist*, 654–663.