Enhancing Abstractive Text Summarization with an Upfront Extractive Layer

Prakash Krishnan

Abstract

Abstractive text summarization of long-form, unstructured documents has been a challenge for researchers. Encoder-only and encoder-decoder transformer models have changed the paradigm and offer many alternatives for an effective summarization. In prior research work, extractive and abstractive summarization were viewed as independent language model tasks. Here in this research paper, we explore a two-step process of abstractive summarization for long-form unstructured documents. As a first step, we conduct an extractive summarization and feed the extractive summarization through an abstractive layer. For the extractive layer, we test our hypothesis by applying different encoders (SBERT and BERT) and classification layers (Clustering, Query Retrieval, and Cosine Similarity). For the abstractive layer, we explore the use of a pre-trained transformer (PEGASUS, Checkpoint-CNN / Daily Mail).

1 Introduction

Abstractive document summarization as a language model task is popular given its application to several use cases such as news summaries, social media reviews, financial summaries, scientific articles, and legal documents. Abstractive summarization is challenging as the input text data can be presented in several formats such as long or short-form text, organized or unorganized, and narrative or conversational style. A single methodology to generate abstractive summarization across all input data types may not work optimally.

Our inspiration for this paper came from the research published by Yang Liu et al., 2019, where a novel document-level encoder based on BERT expressed semantics and generate representations of its sentences. This represented a breakthrough in text summarization by leveraging pre-trained language models as encoders for sentence and paragraph-level documents.

The BERT model suggested by Yang Liu et. al had several challenges for sentence classification such as:

- BERT model was optimized for word or token-level embeddings and not sentence-level embeddings.
- Application of BERT to text summarization is tricky as BERT is trained as a masked-language model and the output vectors are tied to tokens and not sentences.
- To compute accurate sentence similarity with BERT a cross-encoder structure is required where two sentences are passed through a BERT model with a classification head at the top of BERT to derive a similarity score. This is not scalable and causes computational overhead.

Nils Reimers et al., 2019 developed a modified BERT network using siamese network architecture (Sentence-BERT) to derive semantically meaningful sentence embeddings. This opened new tasks such as large-scale semantic comparison, information retrieval, and clustering. These were not possible in the original BERT architecture. See Figure 1 for a Sentence BERT architecture.

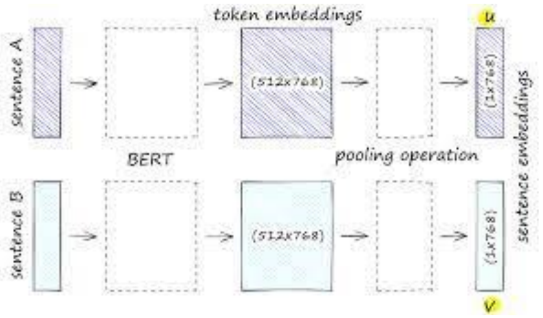


Figure 1: Sentence BERT architecture. Example an input length of 512, an embedding dimension of 768, and a pooling function.

In this research paper, we explore the following novel research questions:

1. For long-form texts such as news articles, would a two-step summarization process with an upfront extractive summarization generated using a sentence transformer and followed by an abstractive summarization lead to superior abstractive summaries as evaluated by ROUGE scores?
2. How can we improve extractive summaries of long documents by leveraging the capabilities of sentence transformers and query retrieval?
3. How do the models perform for different types of datasets such as conversational style MediaSum or more organized CNN/Daily Mail?

2 Background

The novelty of our approach is that prior work mostly treated extractive and abstractive summarizations as separate language model tasks.

Here in this paper, for conversational style text, we explore whether an intermediate enhanced extractive layer with a following abstractive model can lead to superior abstractive summaries.

Models from prior research include the following:

Pre-Transformer Models -
Before transformer models, sentence classification was derived by leveraging topic representation techniques such as Latent Semantic Analysis (Harendra Bhandari et al., 2008), Frequency Driven such as word probability, TF-IDF, and Topic Word approach such as Luhan’s technique.

A Graph-based ranking approach for sentence extraction was developed by Rada Mihalcea (2005, Dept of CS, University of Texas).

Finally on the Machine Learning side, Sequence-to-Sequence RNNs have been proposed by Ramesh Nallapati et al., 2016 for text summarization.

Post-Transformer Models-
In the original BERT-based encoder-decoder model (Yang Liu et al., 2019) the extractive model is built on top of the BERT encoder by stacking inter-sentence transformer layers to capture document-level features. For the abstractive model, an encoder and decoder model with the encoder being the extractive model as described above and the decoder being a 6-layered transformer model.

Sentence-BERT (Nils Reimers et al., 2019) is a novel approach to sentence-level embeddings and text classification using sentence transformers. This is done by mapping each sentence to a vector space and pre-computing sentence vectors that can be stored and then used whenever required. This gave rise to two approaches as illustrated in the paper Sentence-BERT published by Nils Reimers et al., 2019. The first approach was to leverage the standard BERT model and build a sentence embedding by averaging the values across all token embedding outputs or using the first [CLS] token. The second approach is the siamese network architecture (Sentence Level BERT (SBERT)) that derives fixed-size vectors for input sentences. Using a similarity measure like Cosine Similarity, Manhattan, or Euclidean Distance, similar sentences that are semantically close can be ascertained. Both approaches are presented in Section 6 Experimental Results.

3 Dataset

MediaSum data set on Hugging Face is used for model evaluation. This large-scale media interview dataset contains 463.6K transcripts with abstractive summaries, collected from interview transcripts and overview/topic descriptions from NPR and CNN.

For our study, we restricted the dataset size to 1,000 examples for extractive summarization and 100 for abstractive summarization due to computational limitations.

CNN / Daily Mail dataset was also used for comparative purposes given its narrative non-conversational structured nature.

To split the original text into sentences, pre-processing was performed to remove punctuations. (Example: /n, '?</s>, \, _, |, /, Dr., !, ?)

4 Methodology

In this section, we describe our methodology and modeling approach. See Figure 2 for the End-To-End Pipeline.

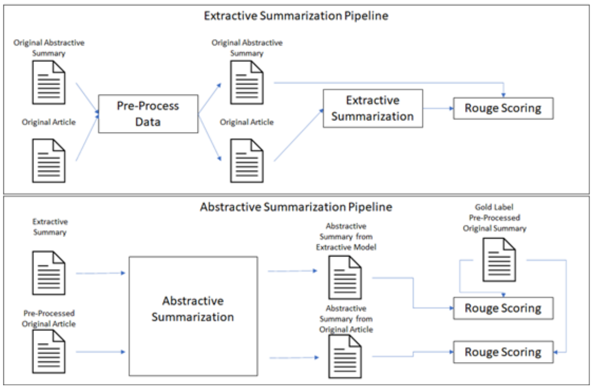

Extractive Summarization Pipeline
Original Abstractive Summary
Original Article
Pre-Process Data
Original Abstractive Summary
Original Article
Extractive Summarization
Rouge Scoring
Abstractive Summarization Pipeline
Extractive Summary
Pre-Processed Original Article
Abstractive Summarization
Abstractive Summary from Extractive Model
Abstractive Summary from Original Article
Gold Label Pre-Processed Original Summary
Rouge Scoring
Rouge Scoring

Figure 2: End-To-End Pipeline. First, sentence summarization is generated using a Classification Layer. Sentence extractions are fed into an encoder-decoder PEGASUS

As a first step, several extractive models were developed using pre-trained encoders BERT and SBERT (Sentence BERT), and a classification layer on top of the encoder. A naive baseline model using TF-IDF was also developed to provide a baseline metric.

Classification techniques explored for the extractive model were:

- Centroid-Sentence Based Summarization where a centroid vector of embedding representation of the input document is derived and the cosine similarity of each sentence vector from the centroid is computed. The Top 'N' sentences with the highest similarity scores were selected for extraction. Top 'N' sentences are computed using a compression ratio that is applied to the original document. The compression ratio is a hyperparameter and user selectable. Trigram blocking is applied to the extraction to ensure words are not repeated in the extractive summary.

- Section-Clustering where the sentences in the input document are clustered using KMeans and the abstractive label is used as a query vector to retrieve sentences from each cluster nearest to the label. The compression ratio is a hyperparameter and user selectable. Trigram blocking is applied to the extraction to ensure words are not repeated in the extractive summary.

The extractive models were run as both unsupervised and supervised models. See Table 1 for a list of extractive models evaluated.

For the extractive summarization, we did not have a gold label extractive summary to assess summarization quality. We used the original abstractive summary as a label to evaluate the different models and classification techniques. We recognize this is not an ideal comparison but offers an intermediate evaluation of the extractive models.

For the supervised extractive model, the input document was first clustered using KMeans, then a query embedding vector was generated using the original article abstractive summary. The distance of the query vector to each cluster centroid was computed. Top N Clusters were derived and then applied KMeans Nearest Neighbor to generate the extractive text summarization. The following hyper-parameters were used - Number of clusters = 5, number of nearest neighbors = 10, and a compression ratio of 0.3.

For the abstractive summarization, we feed the extractive summaries generated into the abstractive model to generate the model summaries. Summaries were generated both from the original full article and also from the extractive summary to determine whether an enhanced extraction provided a superior abstraction. Pre-Trained PEGASUS was used as the abstractive model.

Table 1: Extractive Model Details

Model	Supervised	Selection Criteria
TFIDF	No	Top N Sentences
Pre-trained sentence transformer ('all-MiniLM-L6-v2') with sentence-level embeddings.	No	Agglomerative Clustering
Pre-trained sentence transformer ('all-MiniLM-L6-v2') with sentence-level embeddings.	No	Cosine similarity and degree centrality.

Sentence level built from scratch using the original Bert Model.	No	Cosine similarity and degree centrality.
Pre-trained sentence transformer ('all-MiniLM-L6-v2') with sentence-level embeddings.	Yes	KMeans and Nearest Neighbors Clustering Abstractive Gold Label used as a query

We selected PEGASUS as the abstractive model as it was pre-trained for abstractive summarization. PEGASUS uses an encoder-decoder model for sequence-to-sequence learning. PEGASUS is unique in its pre-training as it factors important sentences in its output closely resembling a human abstractive process. CNN Dailymail checkpoint was used for PEGASUS and no fine-tuning was performed.

5 Evaluation Metrics

For the extractive summarization, we looked at both RL-P and RL-F1 to compare model performances. For this intermediate layer, RL-P was important to ensure the extractive summary captured the key semantics of the abstractive gold summary. However, RL-F1 was a better measure for abstractive summarization.

For the abstractive layer, we compare the model abstractions versus the original abstractive summaries as gold labels.

Additionally, a subjective (manual) evaluation of a random example was used to evaluate for coherence, fluidity, and faithfulness.

6 Experimental Results

The experimental results on MediaSum and CNN/DailyMail datasets are shown in Table 2 and Table 3.

Table 2 presents the experimental results from the extractive models. TF-IDF is the baseline model to compare against.

Table 3 presents the experimental results from the abstractive models. The baseline model is no extractive summarization is performed before abstractive text summarization.

Table 2: Extractive Summarization Results

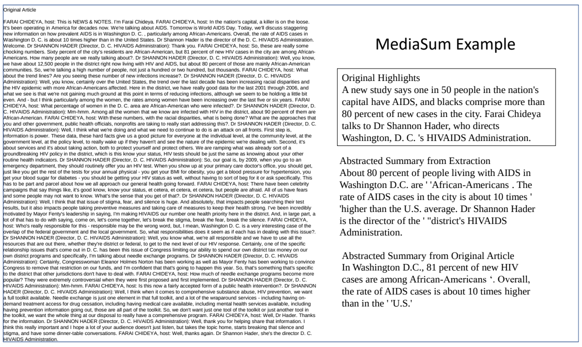
ID	Model	Data	R1	R2	RL-P	RL-F1
Baseline Model						
1.a	TF-IDF	Media Sum	13.71	1.56	15.88	9.21
1.b	TF-IDF	CNN/ DM	15.38	2.96	11.86	11.27
Unsupervised Models						
2.a	SBert Summ	Media Sum	12.69	4.56	46.04	8.82
2.b	SBert Summ	CNN/ DM	23.45	9.00	40.34	16.38
3.a	SBert Trans	Media Sum	12.30	4.49	46.66	8.80
3.b	SBert Trans	CNN/ DM	22.24	8.98	39.78	15.22
4.a	Bert Trans	Media Sum	10.16	3.45	47.04	7.32
4.b	Bert Trans	CNN/ DM	21.03	8.52	41.24	14.17
Supervised Models						
5.a	Supv-SBert	Media Sum	**13.93**	**6.51**	**54.73**	**10.39**
5.b	Supv-SBert	CNN/ DM	**26.58**	**12.92**	**48.70**	**18.65**

Table 3: Abstractive Summarization Results

ID	Model	Data	**Step 1** Ext Layer RL	**Step 2** Abs Layer RL	**Delta**
2.a	SBert Summarizer + PEGASUS	MediaSum	8.82	17.93 22.14**	9.11
2.b	SBert Summarizer + PEGASUS	CNN/DM	16.38	23.10 41.52**	6.72
3.a	SBert Transformer + PEGASUS	MediaSum	8.80	20.35 22.14**	11.55
3.b	SBert Transformer + PEGASUS	CNN/DM	15.22	21.06 41.52**	5.84
4.a	Bert Transformer + PEGASUS	MediaSum	7.32	19.10 22.14**	11.78
4.b	Bert Transformer + PEGASUS	CNN/DM	14.17	23.39 41.52**	9.22
5.a	SBert Transformer + PEGASUS (Supervised)	MediaSum	**10.39**	**23.66** **22.14****	**13.27**
5.b	SBert Transformer + PEGASUS (Supervised)	CNN/DM	**18.65**	**29.17** **41.52****	**10.52**

Note:
** RL-F1 Scores without Extractive Layer

Table 5: Manual Evaluation of Abstractive Summarization for MediaSum



Original Article

FARAI CHIDEYA, host: This is NEWS & NOTES. I'm Farai Chideya. FARAI CHIDEYA, host: In the nation's capital, a killer is on the loose. It's been operating in America for decades now. We're talking about AIDS. Tomorrow is World AIDS Day. Today, we'll discuss staggering new information on how prevalent AIDS is in Washington D. C. , particularly among African-Americans. Overall, the rate of AIDS cases in Washington D. C. is about 10 times higher than in the United States. Dr Shannon Hader is the director of the D. C. HIVAIDS Administration. Welcome. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Thank you. FARAI CHIDEYA, host: So, these are really some chocking numbers. Sixty percent of the city's residents are African-American, but 81 percent of new HIV cases in the city are among African-Americans. How many people are we really talking about?. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Well, you know, we have about 12,500 people in the district right now living with HIV and AIDS, but about 80 percent of those are mainly African-American communities. So, we're talking a high number of people, not just a hundred or two hundred, but thousands. FARAI CHIDEYA, host: What about the trend lines? Are you seeing these number of new infections increase?. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Well, you know, certainly over the United States, the trend over the last decade has been increasing racial disparities and the HIV epidemic with more African-Americans affected. Here in the district, we have really good data for the last 2001 through 2006, and what we see is that we're not gaining much ground at this point in terms of reducing infections, although we seem to be holding a little bit even. And - but I think particularly among the women, the rates among women have been increasing over the last five or six years. FARAI CHIDEYA, host: What percentage of women in the D. C. area are African-American who were infected?. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Mm-hmm. Among all the women that we know are infected with HIV in the district, about 90 percent of them are African-American. FARAI CHIDEYA, host: With these numbers, with the racial disparities, what is being done? What are the approaches that you and other government, public health officials, nonprofits are taking to really start addressing this?. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Well, I think what we're doing and what we need to continue to do is an attack on all fronts. First step is, information is power. These data, these hard facts give us a good picture for everyone at the individual level, at the community level, at the government level, at the policy level, to really wake up if they haven't and see the nature of the epidemic we're dealing with. Second, it's about services and it's about taking action, both to protect yourself and protect others. We are ramping what was already sort of a groundbreaking HIV policy in the district, which is this know your status, HIV tests should be just the same as knowing about your other routine health indicators. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): So, our goal is, by 2009, when you go to an emergency department, they should routinely offer you an HIV test. When you show up at your primary care doctor's office, you should get, just like you get the rest of the tests for your annual physical - you get your BMI for obesity, you get a blood pressure for hypertension, you get your blood sugar for diabetes - you should be getting your HIV status as well, without having to sort of beg for it or ask specifically. This has to be part and parcel about how we all approach our general health going forward. FARAI CHIDEYA, host: There have been celebrity campaigns that say things like, it's good know, know your status, et cetera, et cetera, et cetera, but people are afraid. All of us have fears and some people may not want to know. What's the sense that you get of that?. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Well, I think that that issue of stigma, fear, and silence is huge. And absolutely, that impacts people searching their test results, but it also impacts people taking preventive measures and taking care of measures to keep their health strong. I've been incredibly motivated by Mayor Fenty's leadership in saying, I'm making HIVAIDS our number one health priority here in the district. And, in large part, a lot of that has to do with saying, come on, let's come together, let's break the stigma, break the fear, break the silence. FARAI CHIDEYA, host: Who's really responsible for this - responsible may be the wrong word, but, I mean, Washington D. C. is a very interesting case of the overlap of the federal government and the local government. So, what responsibilities does it seem as if each has in dealing with this issue?. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Well, you know what, we're all responsible and we have to use all the resources that are out there, whether they're district or federal, to get to the next level of our HIV response. Certainly, one of the specific relationship issues that's come out in D. C. has been this issue of Congress limiting our ability to spend our own district tax money on our own district programs and specifically, I'm talking about needle exchange programs. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Certainly, Congresswoman Eleanor Holmes Norton has been working as well as Mayor Fenty has been working to convince Congress to remove that restriction on our funds, and I'm confident that that's going to happen this year. So, that's something that's specific to the district that other jurisdictions don't have to deal with. FARAI CHIDEYA, host: How much of needle exchange programs become more popular? They were extremely controversial when they were first proposed and first implemented. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Mm-hmm. FARAI CHIDEYA, host: Is this now a fairly accepted form of a public health intervention?. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Well, I think when it comes to comprehensive substance abuse, HIV prevention, we want a full toolkit available. Needle exchange is just one element in that full toolkit, and a lot of the wraparound services - including having on-demand treatment access for drug cessation, including having medical care available, including mental health services available, including having prevention information going out, those are all part of the toolkit. So, we don't want just one tool of the toolkit or just another tool in the toolkit, we want the whole thing at our disposal to really have a comprehensive program. FARAI CHIDEYA, host: Well, Dr Hader. Thanks for the information. Dr SHANNON HADER (Director, D. C. HIVAIDS Administration): Well, thank you for helping share that information. I think this really important and I hope a lot of your audience doesn't just listen, but takes the topic home, starts breaking that silence and stigma, and have some dinner-table conversations. FARAI CHIDEYA, host: Well, thanks again. Dr Shannon Hader, she's the director D. C. HIVAIDS Administration.

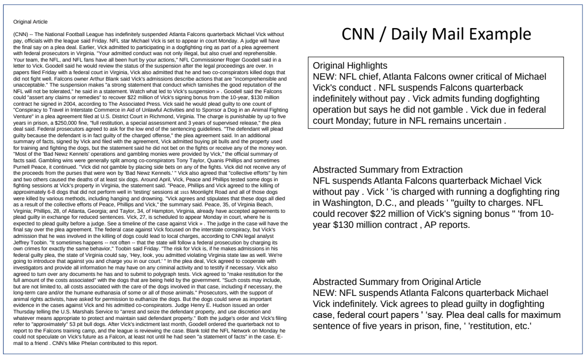
MediaSum Example

Original Highlights
A new study says one in 50 people in the nation's capital have AIDS, and blacks comprise more than 80 percent of new cases in the city. Farai Chideya talks to Dr Shannon Hader, who directs Washington, D. C. 's HIVAIDS Administration.

Abstracted Summary from Extraction
About 80 percent of people living with AIDS in Washington D.C. are ' 'African-Americans . The rate of AIDS cases in the city is about 10 times ' 'higher than the U.S. average. Dr Shannon Hader is the director of the ' "district's HIVAIDS Administration.

Abstracted Summary from Original Article
In Washington D.C., 81 percent of new HIV cases are among African-Americans ‘. Overall, the rate of AIDS cases is about 10 times higher than in the ' 'U.S.'

Table 6: Manual Evaluation of Abstractive Summarization for CNN / Daily Mail



Original Article

(CNN) -- The National Football League has indefinitely suspended Atlanta Falcons quarterback Michael Vick without pay, officials with the league said Friday. NFL star Michael Vick is set to appear in court Monday. A judge will have the final say on a plea deal. Earlier, Vick admitted to participating in a dogfighting ring as part of a plea agreement with federal prosecutors in Virginia. "Your admitted conduct was not only illegal, but also cruel and reprehensible. Your team, the NFL, and NFL fans have all been hurt by your actions," NFL Commissioner Roger Goodell said in a letter to Vick. Goodell said he would review the status of the suspension after the legal proceedings are over. In papers filed Friday with a federal court in Virginia, Vick also admitted that he and two co-conspirators killed dogs that did not fight well. Falcons owner Arthur Blank said Vick's admissions describe actions that are "incomprehensible and unacceptable." The suspension makes "a strong statement that conduct which tarnishes the good reputation of the NFL will not be tolerated," he said in a statement. Watch what led to Vick's suspension » . Goodell said the Falcons could "assert any claims or remedies" to recover $22 million of Vick's signing bonus from the 10-year, $130 million contract he signed in 2004, according to The Associated Press. Vick said he would plead guilty to one count of "Conspiracy to Travel in Interstate Commerce in Aid of Unlawful Activities and to Sponsor a Dog in an Animal Fighting Venture" in a plea agreement filed at U.S. District Court in Richmond, Virginia. The charge is punishable by up to five years in prison, a $250,000 fine, "full restitution, a special assessment and 3 years of supervised release," the plea deal said. Federal prosecutors agreed to ask for the low end of the sentencing guidelines. "The defendant will plead guilty because the defendant is in fact guilty of the charged offense," the plea agreement said. In an additional summary of facts, signed by Vick and filed with the agreement, Vick admitted buying pit bulls and the property used for training and fighting the dogs, but the statement said he did not bet on the fights or receive any of the money won. "Most of the 'Bad Newz Kennels' operations and gambling monies were provided by Vick," the official summary of facts said. Gambling wins were generally split among co-conspirators Tony Taylor, Quanis Phillips and sometimes Purnell Peace, it continued. "Vick did not gamble by placing side bets on any of the fights. Vick did not receive any of the proceeds from the purses that were won by 'Bad Newz Kennels.' " Vick also agreed that "collective efforts" by him and two others caused the deaths of at least six dogs. Around April, Vick, Peace and Phillips tested some dogs in fighting sessions at Vick's property in Virginia, the statement said. "Peace, Phillips and Vick agreed to the killing of approximately 6-8 dogs that did not perform well in 'testing' sessions at 1915 Moonlight Road and all of those dogs were killed by various methods, including hanging and drowning. "Vick agrees and stipulates that these dogs all died as a result of the collective efforts of Peace, Phillips and Vick," the summary said. Peace, 35, of Virginia Beach, Virginia; Phillips, 28, of Atlanta, Georgia; and Taylor, 34, of Hampton, Virginia, already have accepted agreements to plead guilty in exchange for reduced sentences. Vick, 27, is scheduled to appear Monday in court, where he is expected to plead guilty before a judge. See a timeline of the case against Vick » . The judge in the case will have the final say over the plea agreement. The federal case against Vick focused on the interstate conspiracy, but Vick's admission that he was involved in the killing of dogs could lead to local charges, according to CNN legal analyst Jeffrey Toobin. "It sometimes happens -- not often -- that the state will follow a federal prosecution by charging its own crimes for exactly the same behavior," Toobin said Friday. "The risk for Vick is, if he makes admissions in his federal guilty plea, the state of Virginia could say, 'Hey, look, you admitted violating Virginia state law as well. We're going to introduce that against you and charge you in our court.' " In the plea deal, Vick agreed to cooperate with investigators and provide all information he may have on any criminal activity and to testify if necessary. Vick also agreed to turn over any documents he has and to submit to polygraph tests. Vick agreed to "make restitution for the full amount of the costs associated" with the dogs that are being held by the government. "Such costs may include, but are not limited to, all costs associated with the care of the dogs involved in that case, including if necessary, the long-term care and/or the humane euthanasia of some or all of those animals." Prosecutors, with the support of animal rights activists, have asked for permission to euthanize the dogs. But the dogs could serve as important evidence in the cases against Vick and his admitted co-conspirators. Judge Henry E. Hudson issued an order Thursday telling the U.S. Marshals Service to "arrest and seize the defendant property, and use discretion and whatever means appropriate to protect and maintain said defendant property." Both the judge's order and Vick's filing refer to "approximately" 53 pit bull dogs. After Vick's indictment last month, Goodell ordered the quarterback not to report to the Falcons training camp, and the league is reviewing the case. Blank told the NFL Network on Monday he could not speculate on Vick's future as a Falcon, at least not until he had seen "a statement of facts" in the case. E-mail to a friend . CNN's Mike Phelan contributed to this report.

CNN / Daily Mail Example

Original Highlights
NEW: NFL chief, Atlanta Falcons owner critical of Michael Vick's conduct . NFL suspends Falcons quarterback indefinitely without pay . Vick admits funding dogfighting operation but says he did not gamble . Vick due in federal court Monday; future in NFL remains uncertain .

Abstracted Summary from Extraction
NFL suspends Atlanta Falcons quarterback Michael Vick without pay . Vick ' 'is charged with running a dogfighting ring in Washington, D.C., and pleads ' "guilty to charges. NFL could recover $22 million of Vick's signing bonus " 'from 10-year $130 million contract , AP reports.

Abstracted Summary from Original Article
NEW: NFL suspends Atlanta Falcons quarterback Michael Vick indefinitely. Vick agrees to plead guilty in dogfighting case, federal court papers ' ‘say. Plea deal calls for maximum sentence of five years in prison, fine, ' 'restitution, etc.'

7 Discussion and Analysis

We evaluated the extractive and abstractive summarizations using ROUGE. We report unigram and bigram overlap (R1 and R2) and longest common subsequence (RL) for two datasets MediaSum and CNN/DailyMail.

From Table 1 Extractive Summarization Results, the following conclusions can be ascertained:

- TF-IDF baseline model outperformed the unsupervised transformer models for MediaSum on Rl-F1. As MediaSum dataset is conversational style question and answer format, the baseline model gives more importance to words that are frequently mentioned in the question and responses. These frequently mentioned words are also perhaps captured in the abstractive summary resulting in higher RL-F1 scores for the baseline model. (See 1.a, 2.a, 3.a, 4.a)
- We do not see this phenomenon with the CNN/Daily Mail dataset which is more structured, organized, and narrative style. All the

unsupervised models outperformed the baseline model on RL-F1 score. (See 1.b, 2.b, 3.b, 4.b)

- Pre-trained sentence transformer models (SBert Summarizer, SBert Transformer) performed better than the pre-trained word token-based transformers (Bert Transformer). (See 2.a, 2.b, 3.a, 3.b)
- The sentence transformers can capture better sentence-level context than the original BERT model. This was the case with both MediaSum and CNN / Daily Mail.
- A supervised model (SBert Transformer with Kmeans + Nearest Neighbor) leveraging the original abstractive summary as a query for sentence retrievals provided richer extraction as illustrated by the higher ROUGE Scores for both MediaSum and CNN / Daily Mail (See 5.a and 5.b).

From Table 2 Abstractive Summarization Results, the following conclusions can be ascertained:

- For unsupervised models (2.a, 2.b, 3.a, 3.b, 4.a, 4.b) there were improvements in the ROUGE-L scores for both MediaSum and CNN/Daily Mail.
- Extractive summarization seems to provide most benefit when the source document is large and unstructured. MediaSum models were most benefited via the two-step process. CNN/Daily_Mail Models, showed improvements but not as much as MediaSum.
- Extractive summarization generated via query retrieval yielded superior RL scores for the abstractive summarization (5.a, 5.b)
- For all the models in Table 2, stand-alone abstractive summarization yielded superior RL-F1 scores versus the two-step process.
- However, in looking at a random example in Table 5 and Table 6, it does appear that a prelim extractive layer can offer a more complete abstractive summarization as extractive layer can potentially determine underlying semantic similarities.

8 Conclusion and Next Steps

To conclude, there are benefits to performing a preliminary extractive summarization for large unstructured texts leveraging query retrieval where available. Query retrieval can help determine the sentence vectors closest to the query vector and capture semantic similarities. An abstractive layer applied to the extractive summary can yield a coherent and fluent summarization of the original text.

For structured text, an extractive summarization can help but the benefits are lower. Finally, the nature of the input text plays a role in the summarization technique as a conversation-style article would need a dialog summarization model as opposed to narrative summarization. Fine-tuning the model can also be explored to improve ROUGE scores.

9 References

Samir Abdaljalil, and Houda Bouamor. 2021. An Exploration of Automatic Text Summarization of Financial Reports. *In Proceedings of the Third Workshop on Financial Technology and Natural Language Processing.*

Harendra Bhandari, Masashi Shimbo, Takahiko Ito, Yuji Matsumoto. 2008. Generic Text Summarization Using Probabilistic Latent Semantic Indexing. *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I.*

Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders, Yang Liu and Mirella Lapata, *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.*

Yang Liu. 2019. Fine-Tune BERT for Extractive Summarization, *arXiv*.

Jonathan Pilaut, Raymond Li, Sandeep Subramanian, Christopher Pal. 2020. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.*

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.*