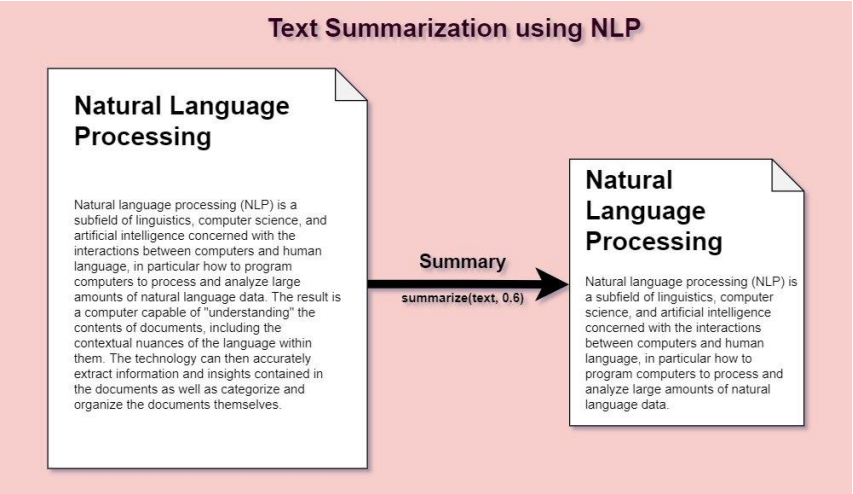
Enhancing Abstractive Text Summarization with an Upfront Extractive Layer

By Prakash Krishnan
November 30, 2022

Berkeley
UNIVERSITY OF CALIFORNIA

Text Summarization

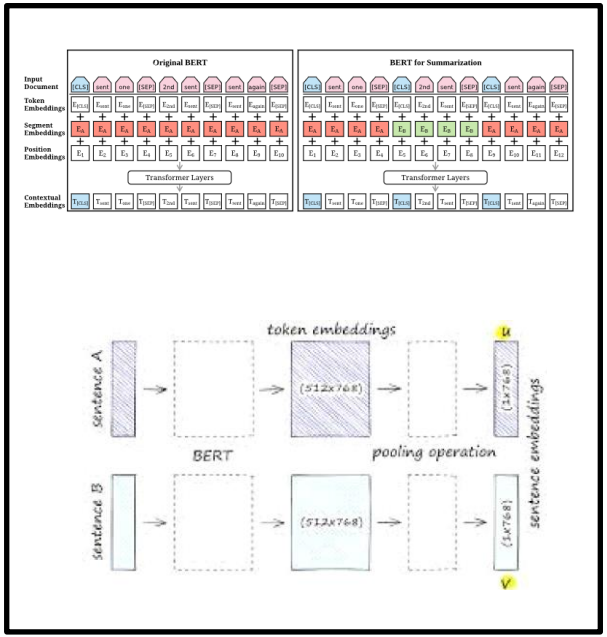
- Language model task of generating a shorter version of a longer document
- Use Cases - News Summaries, Social Media Reviews, Financial Summaries, Scientific Articles, and Legal Documents



Our Inspiration

- BERT Sum Model - Breakthrough research by Yang Liu et al., 2019
  - ❖ BERT was optimized for word or token level embeddings
  - ❖ BERT model not scalable due to cross-encoder structure requiring two sentences passed through the model to compute sentence similarity scores

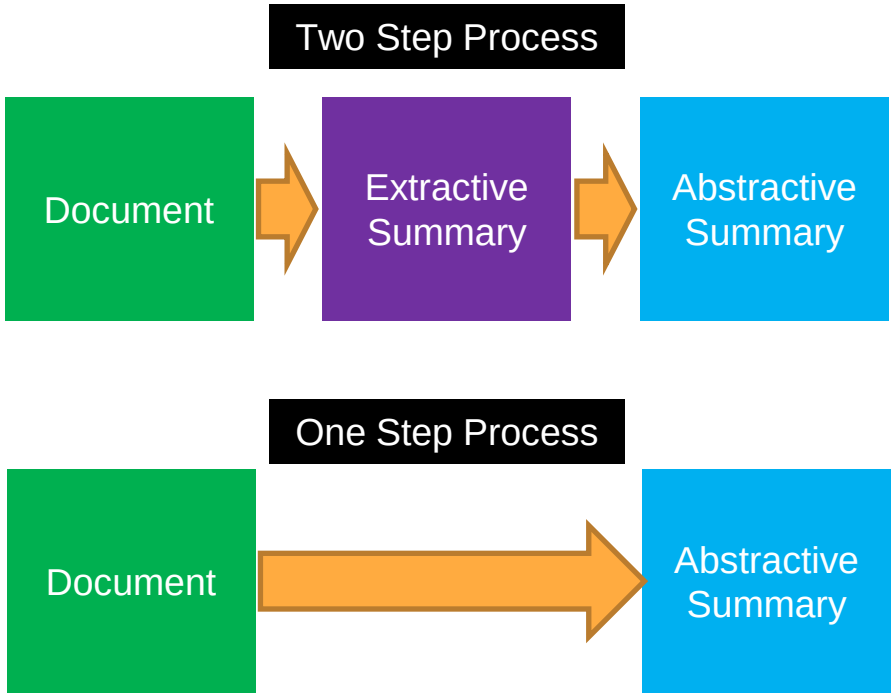
- Sentence BERT Model - Published by Nils Reimers et al., 2019
  - ❖ Siamese Network Architecture that derives sentence level fixed size vectors
  - ❖ Similarity measures (Cosine, Manhattan, Euclidean Distance) can derive sentence vectors that are similar



Berkeley
UNIVERSITY OF CALIFORNIA

Research Questions

1. For long unstructured documents, would a two-step summarization process lead to superior abstractive summaries?
2. How can we improve extractive summaries?
3. How does the nature of input text matter affect model performance?



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

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.

About Dataset

- ✓ MediaSum data set on Hugging Face is used for model evaluation
- ✓ Conversational style dataset that is less structured than CNN / Daily Mail
- ✓ CNN / Daily Mail dataset was used for comparative purposes given its narrative non-conversational nature

Berkeley
UNIVERSITY OF CALIFORNIA

Methodology

Extractive Model

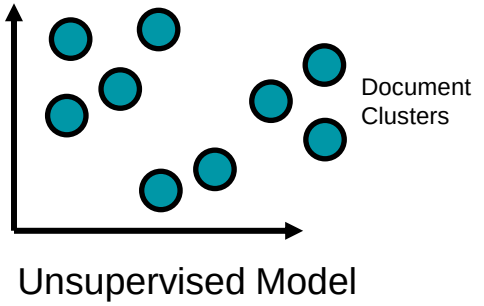
1. Baseline Model – TFIDF
1. Unsupervised versus Self-Supervised Model
1. BERT versus Sentence-BERT

Abstractive Model

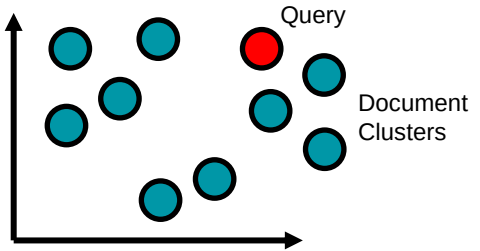
Pegasus Encoder-Decoder Model
(Checkpoint – CNN/Daily Mail)

Extractive Models

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

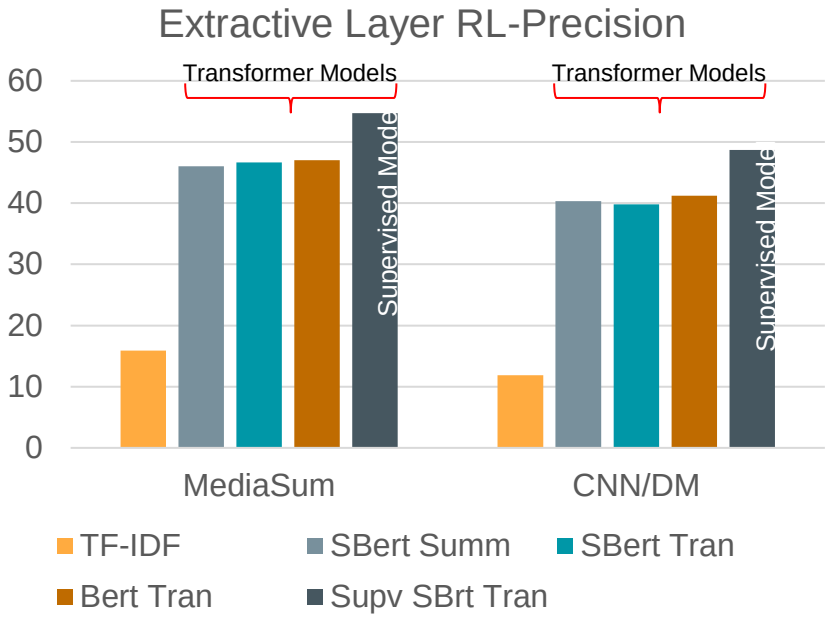


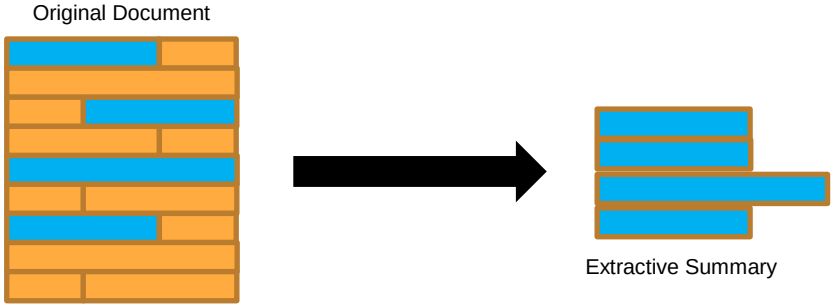
Unsupervised Model



Supervised Model

Experimental Results





Key Takeaways

1. Transformer based models outperformed the baseline TF-IDF

1. Supervised model outperforms the unsupervised models

1. Models performed better on MediaSum versus CNN / Daily Mail

Experimental 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**

Key Takeaways

1. Two step process had the most Rouge-F1 Score improvement for the supervised extractive models
1. MediaSum models showed more promise with a two-step process
1. Supervised model outperforms the unsupervised models
1. Stand-alone abstractive model performed better than the two-step process.
1. Two step on a sample showed superior completeness

Sample Abstraction

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.'

Conclusion

1. Nature of input matters - For long-form documents that are un-organized an upfront extractive layer may result in superior abstraction

1. Supervised extractive layer using query retrieval, paraphrase mining offers superior downstream abstraction

1. Stand-alone abstraction models without an upfront extraction have a role in organized and structured text

1. Sentence transformers capture sentence level embeddings better than token-based transformers

1. Summarizing dialogs needs dialog encoding and summarizing techniques

1. Manual evaluation of fluidity, completeness, coherence and faithfulness