Dan Waters · danwaters@my.unt.edu
University of North Texas
CSCE 5290 · Natural Language Processing
Fall 2021

Abstractive Text Summarization

CSCE 5290 Natural Language Processing | Project Proposal

GitHub: https://github.com/danwaters/nlp-abstractive-text-summarization

Abstract

In a world where everyone is empowered to create text content on a multitude of platforms, how can we quickly and accurately make sense of incomprehensible amounts of information?

We have access to more computing power and performance than any other time in history, and it's a good thing, because it is estimated that 74 Zettabytes of data will be created in 2021[1]. We can reasonably conclude that human society will not be actually reading or analyzing much of it.

Big Data concepts, analytics, cloud computing, and machine learning are commonly applied to reduce the complexity of information and make it digestible for people. A visual dashboard powered by a powerful database engine can turn a million rows of raw data into a handy, intuitive pie chart. Machine learning platforms can intake huge amounts of training data and, over time, structure a model that somehow learned a mathematical representation of a particular domain. The topic of this project is text summarization, and it could help save a tremendous amount of time, especially when coupled with search or other information retrieval techniques.

Use Cases for Text Summarization

Through professional engagements with customers, text summarization has come up more than once as a nice-to-have, and it remains an area of active research. Applications of text summarization include:

- Catching up a user who hasn't paid attention to a chat room (Slack, Teams, Hangouts, etc)
- Summarizing those pesky long corporate emails

[1] Holst, Arne. "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025." *Statista.* Accessed 20 Sep 2021, https://www.statista.com/statistics/871513/worldwide-data-created/

- Creating a timeline overview of related events in law enforcement investigations (for example, "Suspect sends an email in French to Recipient, Recipient opens at 2 AM…")
- Describing trending topics in social media
- Summarizing other common documents such as court proceedings, news articles, social media threads, movie reviews, or any other body of related text.
- Humorous applications such as summarizing song lyrics (what would the summary of a song like Yellow Submarine be?)

Data and Domain

As suggested in the project brief, the Google DeepMind Q&A dataset of CNN and DailyMail articles[2] is a great start for applying the method. Exploratory data analysis and feature engineering will aid in understanding the immediate fit of this dataset for the problem at hand and what transformations, if any, should be executed on the dataset.

As such, this project will be focused on generating summative text for news article content for the express purpose of mastering and comparing various methods. The final domain of text may change, especially if there is time to implement a model that summarizes song lyrics, which will be just for fun.

For even more fun, sentiment analysis could be conducted in addition to summarization. I would like for such a sophisticated model to output silly phrases like *U Can't Touch This is an* ***angry*** *song about* ***hiding items from the audience.***

Proposed Method

Of the two main approaches to summarization (abstractive and extractive), abstractive summarization will be the focus of this paper. There is a certain joy in discovering new phrases and sentences suggested by the model.

There are several machine learning model architectures that are capable of generating text, including recurrent neural networks (RNN), long short-term memory networks (LSTM), autoencoders, and generative adversarial networks (GAN). Having previously experimented with RNN, LSTM, and autoencoder architectures for NLP tasks, I am personally keen to do more work with GAN networks, and it just so happens that there is

[2] Cho, Kyunghun. "DeepMind Q&A Dataset." Accessed 20 Sep 2021, https://cs.nyu.edu/~kcho/DMQA/

a paper[3] to reference on this topic. However, it was authored in 2017, and light years have since elapsed in the world of machine learning research, so I will also be seeking out newer, state-of-the-art processes, using this as a benchmark.

Ethical Considerations

A common pitfall with machine learning is training models on data that leads to bias.

- What if this model always describes articles about a given person or group in unfair terms?
- What if an inaccurate summary leads to harassment of an innocent or unrelated individual? If the model is only trained on the "Opinion" section of these publications, things could get nasty very quickly!

One initial approach is to ensure that the input data is as objectively written as possible, but there is a very real risk of introducing bias into this model and it should be evaluated. Perhaps a neutral sentiment score for generated headlines could be included in the discrimination criteria of the GAN in order to quickly flag output that is too emotionally charged. Feature attribution techniques may also help fine tune the model. Approaches to combat bias will be investigated and applied.

Team

This project will be researched and implemented by Dan Waters individually. The source code will be available at this GitHub repository:

https://github.com/danwaters/nlp-abstractive-text-summarization

This paper can be found here:
https://github.com/danwaters/nlp-abstractive-text-summarization/blob/main/CSCE%205290%20Project%20Proposal%20-%20Draft%20(Waters).pdf

Scope

As this is an individual project which will take some time, expectations, goals, and nice-to-haves are clearly stated here:

[3] Liu, Linqing et al. "Generative Adversarial Network for Abstractive Text Summarization." Accessed 20 Sep 2021, https://arxiv.org/abs/1711.09357

The baseline deliverables consist of a working implementation of a GAN-based abstractive text summarizer trained on the proposed dataset of news articles, detailed analysis of empirical observations, conclusions, future enhancements, and discussion of any improvements or degradation in comparison with the baseline implementation.

Stretch goal

Once an effective model architecture has been identified for the news data, apply the same techniques to a dataset of song lyrics to summarize a song (just for fun). This could also be poems or other literature.

Extra stretch goal

Apply sentiment analysis to describe the emotional quality of the lyrics.

Extra crazy stretch goal

Deploy it as an interactive Twitter bot.

References

1. Holst, Arne. “Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025.” Statista. Accessed 20 Sep 2021, https://www.statista.com/statistics/871513/worldwide-data-created/
2. Cho, Kyunghun. “DeepMind Q&A Dataset.” Accessed 20 Sep 2021, https://cs.nyu.edu/~kcho/DMQA/
3. Liu, Linqing et al. “Generative Adversarial Network for Abstractive Text Summarization.” Accessed 20 Sep 2021, https://arxiv.org/abs/1711.09357