Summary Extraction and Relevance Identification from Spotify Podcasts (Progress Report)

Arijit Ghosh Chowdhury
arijit2@illinois.edu

Dattatreya Mohapatra (Captain)
dm42@illinois.edu

Gargi Balasubramaniam
gargib2@illinois.edu

1 PROBLEM DESCRIPTION

Given the automatic transcript of a podcast episode, the goal is to attempt the following tasks:

(1) Generate meaningful summaries using multiple baseline models and compare them to identify the best performing method
(2) If time permits, identify worst performing summaries and modify baseline models to address their challenges

Our approach involves conducting a qualitative and qualitative evaluation of 3 techniques - TextRank [4], T5 [5] and BERT [6]. We plan to increase complexity in a step by step manner, and observe improvements in the summarization output. To this end, we hope to come up with ways to improve existing baselines after thorough analysis.

2 PROJECT PROGRESS

This section outlines our progress for the project. We performed multiple pre-processing steps to generate data that can be directly fed into our models. We have trained and evaluated two baseline methods – TextRank and T5. Each step is described in detail in the following sections.

2.1 Preliminary Data Analysis

The dataset contains 100,000 episodes from various podcasts across Spotify, sampled between January 1, 2019 and March 1, 2020. [2]. We have also been provided metadata pertaining to each show such as the show URI, the show name, show description, episode duration, to name a few. Table 2 outlines the details of the test dataset on which we have evlaluated our methods.

Table 1: Test Dataset Statistics

Number of Episodes	1000
Average Episode Duration	35.85 Minutes

Table 2: Preliminary Results

Method	ROUGE
TextRank	0.27
T5	0.25

2.2 Methods

We implemented TextRank and T5. Table 2 shows a summary of the quantitative results. For evaluation, we have used the commonly used automatic evaluation metric ROUGE [3].

- **TextRank:** [4] : The algorithm ranks sentences based on their similarity to other sentences. The basic idea implemented by a graph-based ranking model is that when one node links to another one, it is basically voting for that other node. The higher the number of votes that are cast for a node, the higher the relevance of the node.
  This is an **Extractive** summarization method, which outputs sentences already present in the document. ROUGE scores for extractive methods tend to be higher due to higher similarity scores. However, a qualitative evaluation done by human evaluators can help us adjudgde the relevance of abstractive methods better.
- **T5[5]**: This is encoder-decoder model which has been pre-trained on a large corpus of supervised and unsupervised sequence-to-sequence tasks (like translation, summarization, question-answering etc). We have used Huggingface's[1] implementation of the model and ran inference on the first 15 sentences of podcast transcripts.

2.3 Remaining Tasks

T5 surprisingly doesn't perform better that extractive TextRank. This could be because we have used the base model T5 which has been pre-trained on a mixture of tasks. We plant to improve upon it by fine-tuning it by re-training on the podcast transcripts and descriptions.

However, we do not expect fine-tuning to address the main challenges of summarizing podcast data - length of source data, multiple speakers and noisy filler-text. We plan to address the first and last concern by combining TextRank and T5 to perform content selection before training and inference, similar to Zheng et al's [7] approach. The second challenge is a major challenge and will be part of our future scope.

We plan to implement the third method (BERT based representations) and are hoping to incorporate the audio provided as future scope. We will also ask five english speaking volunteers to score the summaries into the defined sprectrum of Bad(B) to Excellent(E).

2.4 Challenges

As a future scope, we hope to leverage the audio samples provided along with the transcript. A key issue with audio is to identify different speakers as there may be multiple speakers in a typical podcast. We plan to explore techniques such as voice separation[1], and then process the audio further.

[1] https://huggingface.co/transformers/model_doc/t5.html

REFERENCES

[1] Shlomo E Chazan, Lior Wolf, Eliya Nachmani, and Yossi Adi. 2021. Single channel voice separation for unknown number of speakers under reverberant and noisy settings. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3730–3734.

[2] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5903–5917. https://www.aclweb.org/anthology/2020.coling-main.519

[3] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[4] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 404–411. https://aclanthology.org/W04-3252

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR* abs/1910.10683 (2019). arXiv:1910.10683 http://arxiv.org/abs/1910.10683

[6] Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-Based Natural Language Generation for Text Summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, 789–797. https://doi.org/10.18653/v1/K19-1074

[7] Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, and Ling Fan. 2020. A Two-Phase Approach for Abstractive Podcast Summarization. *CoRR* abs/2011.08291 (2020). arXiv:2011.08291 https://arxiv.org/abs/2011.08291