

Summary Extraction from Spotify Podcasts

Dattatreya Mohapatra (Captain)
Gargi Balasubramaniam
Arijit Ghosh Chowdhury
Department of Computer Science, UIUC

Motivation



- 1) Podcasts present a genre of datasets where the style of text can vary from very formal to very casual depending on the speakers.
- 2) With the vast amount of data circulating in the digital space, there is need to develop machine learning algorithms that can automatically shorten longer texts and create succinct summaries.

Problem and Contribution



- Given a text transcript from a podcast on Spotify, we aim to generate a few line summary of the podcast.
- We compare against non DL baselines, off the shelf baselines and fine tuned models on the Spotify ASR Dataset.

Dataset




- There are 105360 data points i.e. unique episodes to train our models on.
- On an average, the episode transcript is 5000 words long. Taking an average of 150 words per minute, we get an average podcast length of 33 minutes.

Methods Used



- 1) TextRank
- 2) Text to Text Transfer Transformer (Pre-trained off the shelf)
- 3) Text to Text Transfer Transformer (Fine-Tuned on Spotify Dataset)

Results



The following table outlines the results of the three methods on the validation dataset (subset of processed data). The average of F1 score is taken for calculating the final ROUGE

Method	Rouge (F1)
TextRank	0.109
T5 (Off the shelf)	0.144
T5 (Fine tuned on spotify dataset)	0.222

Insights from Automatic Evaluation (ROUGE)



Method	Rouge (F1)
TextRank	0.109
T5 (Off the shelf)	0.144
T5 (Fine tuned on spotify dataset)	0.312

* While comparing ROUGE scores, we see that the FineTuned model performs well in comparison to other baselines, which confirms our expectation that domain adaptation on the Spotify dataset is a necessary step towards a higher score.

Insights from Manual Evaluation



We used five English speaking volunteers to score the summaries into the defined spectrum of Bad(B) to Excellent(E), as defined by the original paper. This is so that we can effectively capture the subjectivity of how good or bad a summary is, based on how relevant it is to a human evaluator.

- * 3 out of 5 people felt that the summaries generated by TextRank and Fine-Tuned were comparable, and rated it Fair(F).
- * 1 evaluator felt that TextRank is definitely better, and 1 Evaluator felt that given enough data, FineTuned T5 is a much better abstraction of the podcast transcript.
- * 5 out of 5 evaluators agreed that the FineTuned T5 generated better summaries than Off-the-shelf Pretrained T5. This validates our assumption about the need to perform domain adaptation.

Error Analysis and Room For Improvement

- * **Dataset** - Perhaps episode description is not the ideal ground truth to represent summary, since it often contains promotional material which the model learns to generate after every summary, leading to post processing overhead.
- * **Compute** - Since even on the best settings on Colab, the T5 model can only take a certain amount of tokens, perhaps given enough compute T5 has the potential to generate even better summaries. Nevertheless, deep learning based techniques seem to be infeasible for simple use cases.