A Review of a Podcast Text Summarization Approaches from TREC 2020 Proceedings

1 INTRODUCTION

TREC, in collaboration with Spotify, had announced a new Podcasts track in its 2020 edition. There are two tasks in this track - 1) Segment retrieval, and 2) Summarization. In this review, we will focus on the second task and discuss some of the approaches that were submitted by participants.

While there is a lot of literature on text summarization [5, 9, 10, 13] in general, there is no prior precedent for of spoken text summarization. We should note that the methods used for text summarization cannot be directly adapted for podcasts because of the numerous complexities presented by spoken text. Jones et al [3] have very accurately stated that spoken text is very diverse in style, content and format than previously studied speech formats. The style of speaking can be a spectrum from being extremely formal (eg news podcasts) to extremely casual (eg talk show podcasts, movie discussions etc). The dataset provided to us is not representative of any one single topic. The format of podcast can also range from a single speaker to a group discussion. All these dynamics make podcast summarization a more challenging task than general text summarization.

2 DATASET DESCRIPTION

The data distributed by the track organisers consisted of just over 100,000 episodes of English language podcasts. Each episode comes with full audio, a transcript which was automatically generated using Google's Speech-to-Text API as of early 2020, and a description and metadata provided by the podcast creator, along with the RSS feed content for the show.

No ground truth summaries are provided; the closest proxies are the show and episode descriptions provided by the podcast creators. These descriptions vary widely in scope, and are not always intended to act as summaries of the episode content. The organisers have recognised this and filtered the descriptions to establish a subset that is more appropriate as a ground truth set compared to full set of descriptions. The filtering was done with three heuristics – 1) Length (should not be too short or too long), 2) Overlap with show description, and 3) Overlap with description of other shows shown in Table 5. These filters overlap to some extent, and remove about a third of the entire set; the remaining 66,245 descriptions are called the Brass Set.

3 EVALUATION CRITERIA

Due to the subjective nature of summary quality, it is very hard to come up with a single unifying evaluation criteria for judging the correctness and relevance of a summarization models. The problem is compounded for podcast summaries as now we have a large number of possible combinations of genre and format of podcasts (see Section 1) as well. The task organizers have recognised this issue and have presented a three-pronged evaluation method.

(1) Manual assesment and scoring: All submissions will be manually judged by NIST assessors on the Excellent-Good-Fair-Bad (EGFB) scale and will be assigned a score of 4-2-1-0 respectively.
(2) Boolean attributes: Participants are suggested to keep note of a pre-defined set of boolean indicators for good summaries. See Table 1 for more details
(3) ROUGE-L: Generated summaries are compared with the episode descriptions provided by creators and judged by their ROUGE-L [6] scores.

4 BASELINE METHODS

The track organizers have presented findings from five baseline summarization methods for podcasts data:

- onemin: Transcript text for the first one minute of the episode.
- bartcnn: A BART [5] seq2seq model pre-trained on the CNN/Daily Mail corpus for news summarization.
- bartpodcasts: The bartcnn model above, fine-tuned on the brass set described in Section 2
- textranksegments: Chunked the transcript into one-minute chunks, and applied the TextRank algorithm [9]
- textranksentences: The same process as above, except that the transcript is chunked into sentences (instead of 1 min durations) using SpaCy.

The results of each baseline method-episode pair is provided by the organizers as part of the dataset.

5 SUBMITTED APPROACHES

5.1 CUED-SPEECH

Manakul & Gales [7] present a two steps approach: 1) Filtering redundant or less informative sentences in the transcript; 2) Applying a state-of-the-art text summarization system fine-tuned on the Podcast data. They try out different methods of filtering like TextRank [9], truncation, random selection and hierarchical attention (HIER) training [8] and find that the combination of truncation and HIER works the best. After pre-processing the input transcripts, they send it to a modified BART model (described in Section 4). They modify the standard token-level ROUGE loss function into a sequence-level ROUGE-based reward function, inspired from Paulus et al. [11]

5.2 Abstractive Podcast Summarization using BART with Longformer attention

Karlbom & Clifton [4] propose a combined model of. BART and Longformer [1] ("The long document transformer") attention model. Longformer attention scaling addresses the problem of quadratic time complexity in the attention mechanism which many transformer models, including BART, suffer from. The authors combine it with BART by simply replacing its attention mechanism and training the final layers on the podcasts dataset.

Names	Does the summary include names of the main people involved or mentioned in the podcast?
Bio	Does the summary give any additional information about the people mentioned?
Topics	Does the summary include the main topic(s) of the podcast?
Format	Does the summary tell you anything about the format of the podcast?
Title-context	Does the summary give you more context on the title of the podcast?
Redundant	Does the summary contain redundant information?
English	Is the summary written in good English?
Sentence	Are the start and end of the summary good sentence and paragraph start and end points?

Table 1: Boolean indicators of a good podcast summary; meant to guide participants in their method development

5.3 Genre-Aware Abstractive Podcast Summarization

Rezapour et al [12] propose two summarization models that explicitly take genre and named entities into consideration in order to generate summaries appropriate to the style of the podcasts. They argue that the relevancy of a summary varies based on the genre of a podcast. e.g. crime shows might want to use a suspenseful summary; news shows would want to very descriptive. They incorporate this observation into the fine-tuned BART baseline by pre-prending the category metadata of podcasts into the input transcripts.

Their second approach is motivated by a user-survey which indicated that users prefer summaries which mention named-entities. The authors incorporate this by constructing 60-second named-entity-transformed segments from the input transcripts. These segments are then ranked in a TextRank-like fashion and fed into the fine-tuned BART.

5.4 Automatic Summarization of Open-Domain Podcast Episodes

Similar to the previous approaches, Song et al [14] also investigate the the importance of selecting important segments from transcripts to serve as input to state-of-the-art summarizers. They prefer to employ more heuristic-based pre-processing and post-processing approaches in their system. They do content selection by generating saliency-score vectors of each 3-minute segment using combinations of tf-idf scores and duration of participating words.

5.5 A Two-Phase Approach for Abstractive Podcast Summarization

Zheng et al [15] propose a two-phase approach: sentence selection and seq2seq learning. Specifically, they first select important sentences from the noisy long podcast transcripts using a combination of sliding window ROUGE scores and random selection. They further improve this selection using LDA [2] topic modeling to reduce redundancy and increase entropy within topics. semantics. Then the selected sentences are fed into a pre-trained encoder-decoder framework for the summary generation.

REFERENCES

[1] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR* abs/2004.05150 (2020). arXiv:2004.05150 https://arxiv.org/abs/2004.05150

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, null (March 2003), 993–1022.

[3] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2021. TREC 2020 Podcasts Track Overview. *CoRR* abs/2103.15953 (2021). arXiv:2103.15953 https://arxiv.org/abs/2103.15953

[4] Hannes Karlbom and A Clifton. 2020. Abstract Podcast Summarization using BART with Longformer Attention. In *TREC*.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *CoRR* abs/1910.13461 (2019). arXiv:1910.13461 http://arxiv.org/abs/1910.13461

[6] 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

[7] Potsawee Manakul and Mark J. F. Gales. 2020. CUED_speech at TREC 2020 Podcast Summarisation Track. *CoRR* abs/2012.02535 (2020). arXiv:2012.02535 https://arxiv.org/abs/2012.02535

[8] Potsawee Manakul, Mark John Francis Gales, and Linlin Wang. 2020. Abstractive Spoken Document Summarization Using Hierarchical Model with Multi-Stage Attention Diversity Optimization. In *INTERSPEECH*.

[9] 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

[10] Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-Sequence RNNs for Text Summarization. *CoRR* abs/1602.06023 (2016). arXiv:1602.06023 http://arxiv.org/abs/1602.06023

[11] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A Deep Reinforced Model for Abstractive Summarization. *CoRR* abs/1705.04304 (2017). arXiv:1705.04304 http://arxiv.org/abs/1705.04304

[12] Rezvaneh Rezapour, Sravana Reddy, Ann Clifton, and Rosie Jones. 2021. Spotify at TREC 2020: Genre-Aware Abstractive Podcast Summarization. *CoRR* abs/2104.03343 (2021). arXiv:2104.03343 https://arxiv.org/abs/2104.03343

[13] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *CoRR* abs/1704.04368 (2017). arXiv:1704.04368 http://arxiv.org/abs/1704.04368

[14] Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Zhe Feng. 2020. Automatic Summarization of Open-Domain Podcast Episodes. *CoRR* abs/2011.04132 (2020). arXiv:2011.04132 https://arxiv.org/abs/2011.04132

[15] 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