

Abstractive Summarization of **Podcast** Transcripts with **BART** using **Semantic** **Self-segmentation**

Giuseppe Boezio, Giuseppe Murro, Simone Montali

Introduction



Introduction

Podcasts

- Podcasts are a large and growing repository of **spoken audio**.
- Huge *informational upgrade* with respect to news repositories: new topics, higher length, varied styles.
- Summarization is needed by users wanting to **understand what they may like**.



Introduction

Summarization

- Abstractive summarization generates text instead of just picking salient sentences (*extractive summarization*).
- Deep Learning has shown lots of success in this task.
- Spotify published a large-scale dataset containing over 100,000 transcriptions.
- Originally intended for the TREC challenge, the ground truth is just composed by the podcast descriptions.

Literature
review and
past efforts



Literature review and past efforts

BART

- Most of the papers presented during TREC 2020 involved BART pre-trained models
- BART is a *denoising* auto-encoder for pre-training sequence to sequence models



Literature review and past efforts


BART

- Most of the papers presented during TREC 2020 involved

BART pre-trained models

- BART is a *denoising* auto-encoder for pre-training sequence to sequence models

denoising auto-encoder



*Corrupts the text and learns a model
that reconstructs the original*



Literature review and past efforts

BART

- Most of the papers presented during TREC 2020 involved BART pre-trained models
- BART is a *denoising* auto-encoder for pre-training sequence to sequence models
- In addition to BART, several pre-processing techniques differentiated most of the papers



Literature review and past efforts

Related work: challenges

Several challenges:

- Automatic transcriptions with errors
- Disfluencies and redundancies, variety
- Documents are longer than usual



Literature review and past efforts

Related work: solutions

Several ~~challenges~~ solutions:

- Most teams extracted portions of the transcripts to get the most relevant parts: the input exceeds BART-imposed limits (self-attention scales quadratically)
- Manakul et al. used a hierarchical model to filter transcriptions
- Zheng et al. selected sentences using ROUGE
- Song et al. extracted time segments rather than sentences

System description

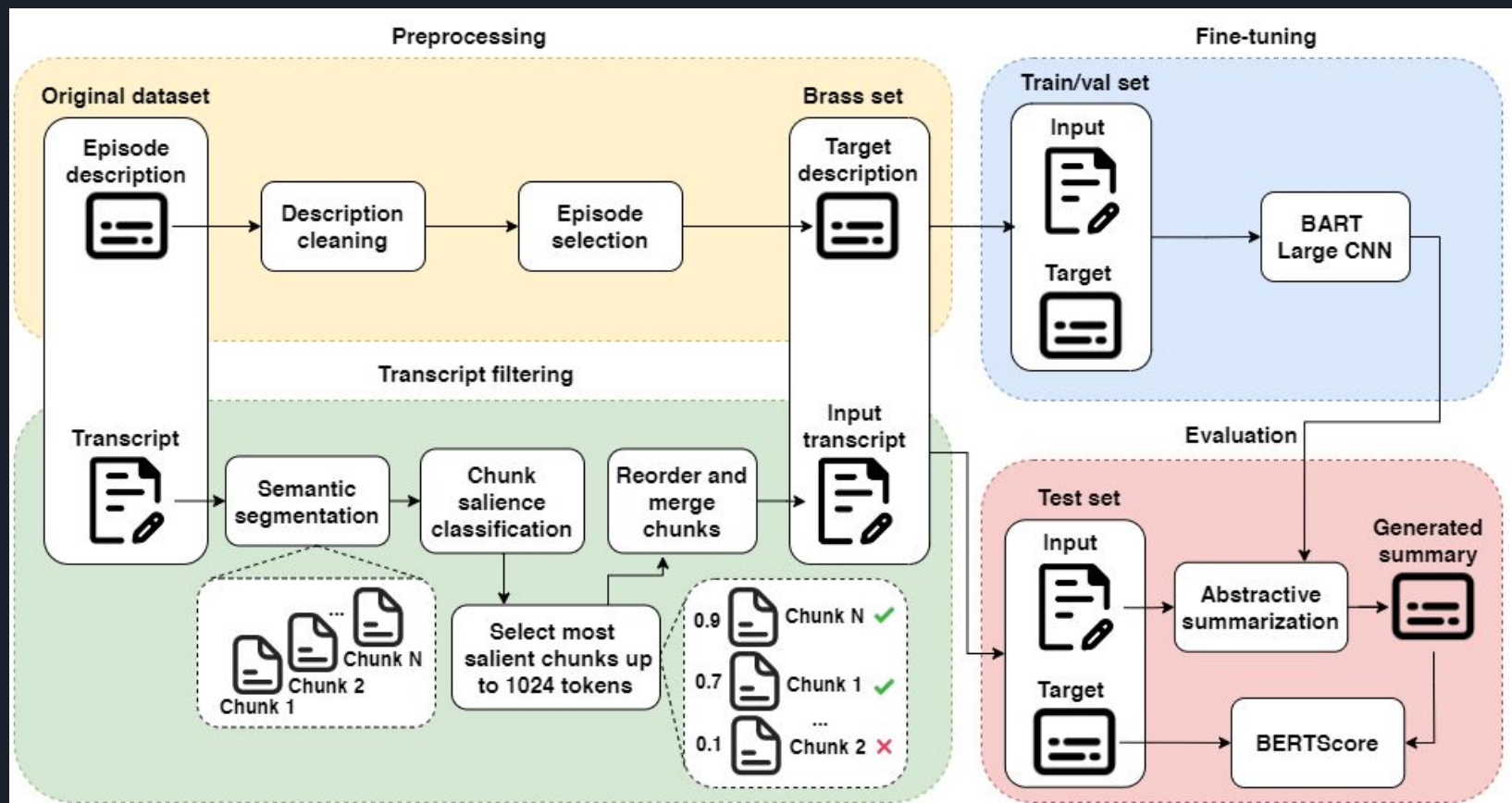


System description

Introduction

Our method is inspired by the past work, combining the strengths.

- Emphasis on the content selection, selecting chunks with an extractive module (classifier-based)
- Extensive pre-processing on descriptions, selecting a subset of the corpus





System description

Preprocessing

Descriptions varied widely in quality, so we cleaned them and removed the ones which were not informative.

- A gold dataset is available (contains good results) and merged
- Removal of improper sentences via heuristics (sponsorship, boilerplate, URLs, mentions, emojis...)
- Removal of non-informative sentences using IDF
- Episode selection on **length, description similarity between show/other episodes**



Examples

Pre-processing

Danielle and Jessi could talk your ears off when it comes to this topic. Episode 004 is all about their skincare routines, products they love, and tips and tricks for feeling radiant and confident in your own skin. Follow them @basicallyorganicpodcast (and @jessimechler @itsdaniellebridges) for tags of all the brands they're currently loving! Rate and subscribe!!

Before

Danielle and Jessi could talk your ears off when it comes to this topic. Episode 004 is all about their skincare routines, products they love, and tips and tricks for feeling radiant and confident in your own skin.

After



Examples

Removed episodes

Life and fashion all packed into a panini

Show description

A real banger, one for the ages

Episode description

Life and fashion all packed into a panini

Episode description

This is a banger, one for the ages

Another episode
description



System description

Semantic segmentation

With podcast transcripts being longer than usual, we need a way of trimming them down.

- Semantic self-segmentation (Se3) computes similarity between sentence embeddings
- Embeddings are computed using Sentence-BERT
- The semantic similarity is then calculated between sentences and averaged by chunk



System description

Chunks filtering

- Salient chunks should be similar to the description, which **at runtime is not available**
- We trained a **classifier** to predict which chunks are therefore **salient**, using the **gold dataset** as training data and computing the ROUGE-L-F1 score with descriptions
- Chunks with score higher than a threshold ($\tau = 0.2$) are labeled as positive and kept in the **brass set**.

Experimental setup



Experimental setup

Fine tuning

- We fine-tuned bart-large-cnn, pre-trained on CNN Daily Mail
- The training has been performed in the cloud, with 4x NVIDIA RTX3090 GPU (24GB) and 128GB of RAM
- The train/dev split contains approx. 70k/7.7k examples
- Optimization is performed on BART's loss.
- The model has finally been uploaded to HuggingFace hub.



Experimental setup

Evaluation

- The final test set consists of 1025 episodes (+2 removed)
- BART hyperparameters were selected following state-of-the-art

length_penalty=2.0 num_beams=4 no_repeat_ngram_size=3 max_length=250 min_length=39

- Two metrics were chosen: ROUGE and BERTscore
 - ROUGE is meaningful but syntactic, matching n-grams (suboptimal in abstractive summarization)
 - BERTscore is semantic, using a contextual word embedding

Results



Results

Metrics

Both ROUGE and BERTscore were computed using precision, recall and F1.

- ROUGE was computed in the -1 (unigram), -2 (bigram), -L (longest common subsequence) and -L-Sum (LCS w/ summary) versions.
- BERTscore was computed both in the standard and IDF weighted version.

Results

Metric	Model	Precision	Recall	F1 Score
ROUGE-1	bart-large-cnn	0.2357	0.2125	0.1917
	bart-large-finetuned	0.3250	0.2315	0.2370
ROUGE-2	bart-large-cnn	0.0449	0.0447	0.0379
	bart-large-finetuned	0.0884	0.0677	0.0670
ROUGE-L	bart-large-cnn	0.1467	0.1408	0.1223
	bart-large-finetuned	0.2160	0.1634	0.1621
ROUGE-L Sum	bart-large-cnn	0.2088	0.1855	0.1684
	bart-large-finetuned	0.2846	0.2003	0.2058

Model	Precision		Recall		F1 Score	
IDF weighting	Yes	No	Yes	No	Yes	No
bart-large-cnn	0.8103	0.8317	0.7941	0.8121	0.8018	0.8214
bart-large-finetuned	0.8401	0.8631	0.8093	0.8279	0.8240	0.8447



Results

Analysis

- Fine-tuning noticeably improved the model.
- ROUGE-2 is heavily lower than the others, possibly because the same bi-grams are difficult to find in abstractive summaries.
- BERTscore showed that the meanings are often correct, reaching 82.40% F1 with IDF weighting, increasing by more than 2% from the baseline.
- Transcript filtering was proven crucial for the performance.
- A longer training might have been useful.



Examples

Summaries

It should be an expected thing.
Why is there this perception that like, oh, yeah, you're a woman you have to have kids once you get married.
I think I'm still figuring, you know myself out and where I'm going to go with everything but I've definitely got a better idea.
Sometimes it's hard to not want to always go with the flow.

`bart-large-cnn`

In this episode we talk about women in the sport of fishing. We talk about what it's like to be a female angler and how it's different than being a male angler, and how to deal with the expectations that come with being a woman angler.

`bart-large-finetuned`



Results

Conclusions and future work

- Extractive models have now been mostly outdated by abstractive ones.
- It is challenging to generate accurate summaries, even for humans.
- A next step could be an ablation study to understand whether the transcript filtering actually improved performances, re-training the model without this component.
- Entity recognition and discourse analysis could be investigated in the future to extract semantic information for the search of salient parts in the transcript.

Thanks!